



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

Doctoral Dissertation

Optimal Power Delivery Strategy in
Modern VLSI Design

Seungwon Kim

Department of Electrical Engineering

Graduate School of UNIST

2019

Optimal Power Delivery Strategy in Modern VLSI Design

Seungwon Kim

Department of Electrical Engineering

Graduate School of UNIST


Optimal Power Delivery Strategy in Modern VLSI Design

A dissertation
submitted to the Graduate School of UNIST
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Seungwon Kim

6 / 18 / 2019

Approved by



Advisor

Kyung Rok Kim

Optimal Power Delivery Strategy in Modern VLSI Design

Seungwon Kim

This certifies that the dissertation of Seungwon Kim is approved.

6 / 18 / 2019

signature



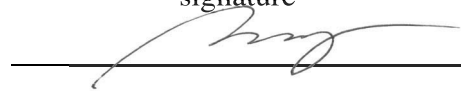
Advisor: Kyung Rok Kim

signature



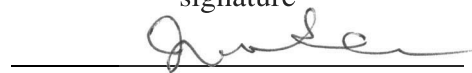
Co-Advisor: Seokhyeong Kang

signature



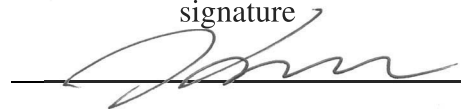
Seong-Jin Kim

signature



Jongeun Lee

signature



Youngmin Kim

DEDICATION

- *To my parents, **Sookyoung Kim** and **Younghi Jeon**, and my sister **Jiyeon Kim**, without their endless love and encouragement this dissertation would not have been finished.*

Abstract

In a modern very-large-scale integration (VLSI) designs, heterogeneous architectural structures and various three-dimensional (3D) integration methods have been used in a hybrid manner. Recently, the industry has combined 3D VLSI technology with the heterogeneous technology of modern VLSI called *chiplet*. The 3D heterogeneous architectural structure is growing attention because it reduces costs and time-to-market by increasing manufacturing yield with high integration rate and modularization. However, a main design concern of heterogeneous 3D architectural structure is power management for lowering power consumption with maintaining the required power integrity from IR drop. Although the low-power design can be realized in front-end-of-line level by reduced power supply complementary metal–oxide–semiconductor technologies, the overall low-power system performance is available with a proper design of power delivery network (PDN) for chip-level modules and system-level architectural structure. Thus, there is a demand for both the coanalysis and optimization for both chip-level and system-level. We analyzed and optimized power delivery on-chip in various 3D integration environments, and we also have proposed a chip-package-PCB coanalysis methodology at the system level. For through-silicon-via (TSV)-based 3D integration circuit (IC), We have investigated and analyzed the voltage noise in a multi-layer 3D stacking with partial element equivalent circuit (PEEC)-based on-chip PDN and frequency-dependent TSV models. We also have proposed a wire-added multi-paired on-chip PDN structure to reduce voltage noise to reduce IR drop. The performance of TSV-based 3D ICs has also been improved by reducing wake-up time through our proposed adaptive power gating strategy with tapered TSVs. For die-to-wafer 3D IC, we have proposed a power delivery pathfinding methodology, which seeks to identify a nearly optimal PDN for a given design and PDN specification. Our pathfinding methodology exploits models for routability and worst IR drop, which helps reducing iterations between PDN design and circuit design in 3D IC implementation. We also have extended the observation to system-level, we have proposed a power integrity coanalysis methodology for multiple power domains in high-frequency memory systems. Our coanalysis methodology can analyze the tendencies in power integrity by using parametric methods with consideration of package-on-package integration. We have proved that our methodology can predict similar peak-to-peak ripple voltages that are comparable with the realistic simulations of high-speed low-power memory interfaces. Finally, we have proposed analysis and optimization methodologies that are generally applicable to various integration methods used in modern VLSI designs as computer-aided-design-based solutions.

Contents

I. Introduction	1
1.1 Difficulties in Power Delivery in Modern VLSI Design	1
1.2 This Dissertation	5
II. Background and Prior Work of Power Delivery Optimization in Modern VLSI Designs	8
2.1 Power Gating	8
2.2 Power Delivery Network in On-Chip	10
2.3 System-level Power Delivery Analysis	12
III. Adaptive Power Gating Strategy and Tapered TSV in Multi-Layer 3D IC	14
3.1 Preliminary Analysis of Power Delivery Network	16
3.1.1 Analysis of power delivery network in the TSV-based 3D IC	16
3.1.2 Multi-paired PDN structure for reduction of voltage noise	19
3.2 TSV-Based 3D IC Power Delivery Network	23
3.2.1 TSV EM modeling and S-parameter extraction	23
3.2.2 PEEC-based on-chip PDN	25
3.3 Power Gating Design in 3D IC	27
3.3.1 Power gating	28
3.3.2 Single-stage power gating with daisy-chain buffer	28
3.4 Adaptive Two-Stage Power Gating Strategy in a 3D IC	30
3.5 Optimization of Adaptive Interval	34
3.6 Effect of Tapered TSV Structure	37
3.7 Conclusions and Future Directions	43
3.8 Acknowledgments	44
IV. Power Delivery Pathfinding for Emerging Die-to-Wafer Integration Technology	45
4.1 Related Work	46
4.2 Methodology	48
4.2.1 Power delivery pathfinding flow	48
4.2.2 PDN design knobs	48
4.2.3 WIR & routability modeling	49

4.3	Experiments	53
4.3.1	Scalability study	53
4.3.2	Sensitivity study	54
4.3.3	IR drop model	57
4.3.4	Routability model	58
4.3.5	Verification of pathfinding on real design	60
4.4	Conclusions and Future Directions	62
4.5	Acknowledgments	62

V. Power Integrity Coanalysis Methodology for Multi-Domain High-Speed Memory Systems

		64
5.1	Power Delivery System Analysis Model	66
5.1.1	Multi-domain power distribution network	66
5.1.2	Input power source	67
5.1.3	On-chip modeling	69
5.1.4	Decoupling capacitor, wire and ball modeling	71
5.2	Procedure of Proposed Methodology	74
5.3	Simulation and Analysis	76
5.3.1	Preliminary analysis of power domain coupling	76
5.3.2	Model verification	81
5.3.3	Domain coupling	82
5.3.4	On-chip decap effect	82
5.3.5	Input noise effect	85
5.4	Conclusions and Future Directions	87
5.5	Acknowledgments	87

VI. Conclusion and Future Consideration

Bibliography

List of Figures

Figure 1.1:	Changes in the integrated number of transistors per unit cost with the technology nodes in the VLSI chip. The costs beyond 20 nm node are predicted values [93].	2
Figure 1.2:	Growing gap between transistor delay and interconnect delay in advanced technology nodes [2].	2
Figure 1.3:	Multichip architecture revolution from monolithic die to <i>chiplet</i> -based system [6]	3
Figure 1.4:	(a) An example of 3D <i>chiplet</i> architectural structure [94], and (b) side view of hybrid heterogeneous integration of 3D system-in-package and 3D system-on-chip [95].	4
Figure 1.5:	The on-chip routability versus IR drop envelope in high design solution space of on-chip PDN.	5
Figure 1.6:	Scope of this dissertation.	6
Figure 2.1:	(a) Power gating operation with (b) PMOS header switch [79].	9
Figure 2.2:	Examples of 3D IC integration technologies. (a) wafer-on-wafer integration, and (b) die-to-wafer integration.	12
Figure 2.3:	An example of a PDN design of one die in a 3D IC containing VIs.	12
Figure 3.1:	TSV-based face-to-back bonding of the multi-layer 3D IC.	15
Figure 3.2:	Power and ground TSV structure with six ports for frequency-dependent <i>S</i> -parameter generation [32].	17
Figure 3.3:	3D view of (a) the regular and (b) the multi-paired PDN structure. (c) Top view and corresponding VDD and GND ports in M5 and M6. PDN is generated using M6 and M5 with 3×3 VDDs and GNDs. Green line is VDD and blue line is GND metal, respectively.	18
Figure 3.4:	(a) VDD droop (top) and GND bump (bottom) comparison with TSV and without TSV in a single layer of die. (b) Impact of the number of layers and the inductance (<i>L</i>) on the PDN voltage noise.	20
Figure 3.5:	Top view of the wire-added multi-paired (two pairs) PDN structure of (a) 20 μm (minimum) wire space, and (b) 120 μm wire pair space.	21
Figure 3.6:	Top view of two VDD (green) and GND (blue) wires are added to the proposed multi-paired (three pairs) PDN structure.	22
Figure 3.7:	The worst VDD droop comparison in the conventional regular (non-paired) PDN and our various proposed multi-paired PDN structure.	23

Figure 3.8:	Insertion loss (top) and return and coupling loss (bottom) in S -parameter of the power TSVs as the frequency increases. The geometry of the 18 TSVs (9 VDD and 9 GND) is shown in the top figure, including port numbers.	24
Figure 3.9:	3D IC stacking for a total of five layers with power gating switches and logic in each layer. Face-to-back stacking is used. VDD and GND are supplied from the bottom C4 bumps. The proposed wake-up controller minimizes the wake-up latency of each layer; it controls the interval delay between the <i>enable_few</i> and <i>enable_rest</i> signals.	26
Figure 3.10:	VDD drop (a) and GND bump (b) comparison with and without a TSV in a single-layer die. A clock buffer that is 40 times larger than the minimum clock buffer with an input slew of 100 ps is used in this simulation.	26
Figure 3.11:	On-chip PDN model of one layer of a 3D IC with PMOS header switches for power gating. The enable-low control signal propagates through all of the switches, with additional buffers in the daisy chain.	27
Figure 3.12:	(a) Conventional one-stage turn-on scheme with a daisy chain, (b) two-stage scheme using the <i>enable_few</i> and <i>enable_rest</i> signals, and (c) proposed adaptive two-stage turn-on scheme. The interval between the <i>enable_few</i> and <i>enable_rest</i> signals is controlled by the wake-up controller (see Figure 3.9) on the basis of victim- and aggressor-layer information.	29
Figure 3.13:	The <i>enable_few</i> (black line) and <i>enable_rest</i> (red line) signals for different few-to-rest ratios. The VVDD waveform is shown by the blue line. The few-to-rest ratios are shown at the bottom of each plot. “Few 1st” means the first turned on <i>enable_few</i> switch; “Few last” is the last switch. The ratios are (a) 0.25:9.75, (b) 0.5:9.5, (c) 1:9, and (d) 2:8.	32
Figure 3.14:	IR-drop dependency according to the location of the aggressor and victim layers. The IR drop in the victim layer increases when the aggressors are located in adjacent layers. The absolute values are summarized in the table included in the figure.	33
Figure 3.15:	Voltage drop profiles of the three combinations. The blue line is the virtual VDD of the worst victim layer. The vertical black and red lines represent the starting time of the <i>enable_few</i> signal and the <i>enable_rest</i> signal, respectively. The vertical dotted line is the worst wake-up time among aggressors. (a) The best IR-drop combination, which has a 0 ns interval between the <i>enable_few</i> and <i>enable_rest</i> signals. (b) The case of a 6.5 ns interval. (c) The worst-case IR-drop combination, which requires a 12.5 ns interval.	35
Figure 3.16:	Adaptive wake-up time profile for each aggressor layer based on the layer configurations, IR-drop requirement of the victim layer, and idle layers in the 3D IC. A total of 180 cases of layer combinations for the wake-up situation. Overall average wake-up time (of each aggressor layer) is reduced because it considers the idle state, which is ignored in the simulations in Section 3.4.	37

Figure 3.17: Uniform TSV (left) and tapered TSV (right) examples. Larger TSVs satisfy larger current demands. Therefore, we place larger TSVs near the VDD source (i.e., lower layers) and smaller TSVs in the upper layers.	38
Figure 3.18: Frequency-dependent extracted RLGC [(a) resistance, (b) inductance, (c) conductance, and (d) capacitance] values of TSVs in different diameters.	39
Figure 3.19: Wake-up time profile with tapered TSV cases 1, 2, and 3. The black dot line is average wake-up time of conventional 2-stage method and the red dot line is average wake-up time of conventional 2-stage method with tapered TSV only. The results show not only an improved wake-up time but also more balanced values among layer configurations (i.e., victim, aggressor, and idle) by the tapered TSV than the uniform TSV.	41
Figure 3.20: Normalized bar graph of the wake-up time when using the proposed method and KOZ with various tapering cases.	42
Figure 4.1: Model-based PDN pathfinding flow which gives the optimal PDN design considering both IR drop requirement and routability requirement.	49
Figure 4.2: Illustration of circuit design-independent PDN design knobs.	49
Figure 4.3: (a) WIR modeling flow. (b) Routability modeling flow.	51
Figure 4.4: Illustration of (a) mesh-like placement as in [45], and (b) our 3D mesh-like placement with VIs.	51
Figure 4.5: Routability (K_{th}) versus #inst reflecting various number of PDNs.	54
Figure 4.6: WIR (left) and routability (right) sensitivity results for circuit-independent knobs width (top) and set-to-set pitch (bottom). The red numbers indicate the slope of the K_{th} change with each knob.	55
Figure 4.7: WIR (left) and routability (right) sensitivity analysis results for circuit-independent knobs width (top) and set-to-set pitch (bottom) with various utilization.	56
Figure 4.8: Routability sensitivity analysis results for circuit-independent knobs (a) width and (b) set-to-set pitch with various VI densities.	57
Figure 4.9: Results for WIR model.	58
Figure 4.10: Correlation of routability graph between actual K_{th} and predicted K_{th} by each knob-based (Model1), routing capacity score-based (Model2) and combined (Model3) models. The scatter points displayed in the graph represent a total of 88 #testing points and a total of 168 #PDNs training points.	59
Figure 4.11: Correlation of routability graph between actual K_{th} and predicted K_{th} values by each of (a), (b) knob-based (Model1) and (c), (d) combined (Model3). The scatter points displayed in the graph represent a total of 256 #testing points and a total of 256 #PDNs training points.	60

Figure 4.12: Routability (K_{th}) versus IR drop data with PDN design knobs for (a) AES encryption core and (b) JPEG encoder testcases. Blue dots denote trained ranking of PDNs and are represented by the second y-axis as K_{th} values. Optimal, reference and worst PDNs are verified by real designs. The red arrows indicate improvement from the reference PDN. The red region indicates WIR greater than the WIR drop of the reference PDN.	63
Figure 5.1: Process of our proposed PI coanalysis methodology.	65
Figure 5.2: Overview of the PDS for a PoP model.	66
Figure 5.3: Structure and parameters of a sample PDN: (a) slant view, (b) side view, and (c) top view.	68
Figure 5.4: (a) Four-element linear VRM model and (b) the proposed two-element linear input power source model with a sinusoidal source as the noise.	69
Figure 5.5: Lumped RLC T-model.	70
Figure 5.6: (a) Real on-chip current profile on the pad of VDDQ. (b) Overview of the characterized pseudo-random current profile parameters.	71
Figure 5.7: Procedure of our proposed fast analysis methodology for PI.	74
Figure 5.8: Schematic of PoP PDS model and measurement point for on-chip PI analysis. ..	77
Figure 5.9: Locations of twenty ports for the simulations of PDN.	78
Figure 5.10: Z-parameters of a 20-port PDN structure: CGND = 1, RGND = 1, MARGIN_WIDTH = 2000 μm , CORE_THICKNESS = 400 μm , PSR_THICKNESS = 200 μm	78
Figure 5.11: 2-port network system of PDN consist of capacitors.	79
Figure 5.12: PDN structures according to existence and nonexistence of ground margin: (a) structure of PDN that has no ground margin (MARGIN_WIDTH = 0 μm), (b) structure of PDN that has ground margin (MARGIN_WIDTH = 2000 μm).	80
Figure 5.13: PDN structure according to existence and nonexistence of ring ground plane and center ground plane: (a) structure of PDN that has both ring and center ground plane (RGND = 1 and CGND = 1), (b) structure of PDN that has center ground plane and no ring ground plane (RGND = 0 and CGND = 1), (c) structure of PDN that has ring ground plane and no center ground plane (RGND = 1 and CGND = 0).	81
Figure 5.14: Transient results of (a) generated current profile and corresponding voltage fluctuations in two methods (b) voltage with three PDN structures (with both center/ring, without ring, without the central ground) and (c) three ground margins (0 μm , 1500 μm , 3000 μm) of the package PDN on the pad of the on-chip VDDQ.	83
Figure 5.15: Current profile (above) and voltage graph (below) on the pad of the on-chip VDDQ. IR drop on the on-chip pad decreases as the on-chip decap increases. ...	84
Figure 5.16: IR drop (left) and resonance frequency (right) on the pad of on-chip with various values of the on-chip decap	84
Figure 5.17: Transient result of the voltage (above) and the current profile (below) on the pad of the on-chip VDDQ.	86

Figure 5.18: IR drop on the on-chip pad with various input noise frequencies; (a) sweep simulations with 100 MHz steps, (b) Monte Carlo simulations.	86
Figure 6.1: High-level overall flow for identification of best PDN.	92
Figure 6.2: An example of a graph of the possible PDN combination cases for a total of nine BEOL layers. Metal layer 2 is assumed to be a power rail.	93

List of Tables

Table 3.1:	Structural parameters of the metal interconnect and TSV.	17
Table 3.2:	The worst VDD droop in the conventional regular PDN and the proposed multi-paired PDN.	19
Table 3.3:	The worst VDD droop in the wire-added multi-paired (two pairs) PDN structure. 80 μm space is impossible, because the GND line runs over the VDD ports in M6 layer.	21
Table 3.4:	Worst-case IR drop and wake-up times for different few-to-rest ratios	31
Table 3.5:	Conventional and proposed wake-up times with different tapering cases	38
Table 4.1:	PDN design knobs.	50
Table 4.2:	Reference design of PDN.	53
Table 4.3:	Sensitivity to VI densities (#nets = 25172).	57
Table 4.4:	Summary of design solution space.	58
Table 4.5:	Information of testcases.	61
Table 4.6:	Simulation results with real testcases for AES cipher and JPEG encoder. All K_{th} values are average values with five de-noising runs.	61
Table 5.1:	Modeling parameters and dimensions of multi-domain PDN. The value is the parameter value that is the reference of the PDN structure to be used in the simulation.	88
Table 5.2:	Sinusoidal source parameters.	88
Table 5.3:	Values of the used parameters that significantly affect the simulation results.	89
Table 5.4:	Margin effect about the power domain coupling.	89
Table 5.5:	Ground plane effect about the power domain coupling	89
Table 5.6:	PSR_THICKNESS and CORE_THICKNESS effect about the power domain coupling	90
Table 5.7:	Comparison of peak-to-peak ripple voltage results obtained by full layout <i>SPICE</i> simulation and proposed methodology.	90

ACKNOWLEDGMENT

I would like to thank the numerous people I have worked with in the process of completing this work, in the lab and beyond.

First and foremost, I would like to express the deepest appreciation to my academic advisor, Professor Seokhyeong Kang, for the continuous support of my Ph.D study and related research, for his patience, motivation, and immense knowledge. Besides constant encouragement, support and guidance on my research, he also provided me a lot of opportunities to meet with other leading experts from academia and industry. Without his guidance and persistent help this dissertation would not have been possible.

I am also greatly indebted to advisor Professor Kyung Rok Kim for the many inspiring discussions we had and his fresh and sharp perspective on my work. His expertise allowed us to stand on a strong foundation and get to the solutions much more quickly than we would have on our own.

I would like to thank my dissertation committee members, Prof. Seong-Jin Kim, Prof. Jongeun Lee and Prof. Youngmin Kim for taking time out of their busy schedules to review and evaluate my research work. I am grateful for their valuable feedback. In particular, I am grateful to Professor Youngmin Kim for enlightening me the first glance of research.

My sincere thanks also goes to Prof. Ki Jin Han, Prof. Andrew B. Kahng, who provided me an opportunity to collaborative research. Without they precious support and valuable advice, it would not be possible to conduct this research.

I also wish to thank my labmates in CAD & SoC Design Lab, Dr. Yesung Kang, Sunmean Kim, Daeyeon Kim, Yoonho Park, Sunghoon Kim, Sungyun Lee, Sunghye Park, Eunji Kwon, Taeho Lim, Jaewoo Kim, Mingyu Woo, Sanggi Do and Jaemin Lee. We had wonderful projects, and we spent great times together. I wish them the best in their future.

Last but not least, I would like to express my deepest gratitude to my parents, Sookyoung Kim and Younghi Jeon, and my sister Jiyeon Kim for their encouragement throughout my years of study and through the process of researching. This dissertation would not have been possible without their warm love, continued patience, and endless support. I am also thank to my friends, Dr. Sangho Ha, Saebyuk Shin and Sooho Chang, who have supported me along the way.

The material in this dissertation is based on the following publications.

- Chapter III is based on:

- **Seungwon Kim** and Youngmin Kim, “Analysis and Reduction of the Voltage Noise of Multi-layer 3D IC with Multi-paired Power Delivery Network”, *IEICE Electronics Express* 14(18) (2017), pp. 1-9.
- **Seungwon Kim**, Seokhyeong Kang, Ki Jin Han and Youngmin Kim, “Novel Adaptive Power Gating Strategy and Tapered TSV Structure in Multi-layer 3D IC”, *ACM Transactions on Design Automation of Electronic Systems* 21(3) (2016), 44.
- **Seungwon Kim**, Seokhyeong Kang, Ki Jin Han and Youngmin Kim, “Novel Adaptive Power Gating Strategy of TSV-based Multi-layer 3D IC”, *Proc. IEEE International Symposium on Quality Electronic Design*, 2015, pp. 537-541.
- **Seungwon Kim**, Ki Jin Han, Seokhyeong Kang and Youngmin Kim, “Analysis and Reduction of Voltage Noise of Multi-layer 3D IC with PEEC-based PDN and Frequency-dependent TSV Models”, *Proc. IEEE International SoC Design Conference*, 2014, pp. 124-125.

- Chapter IV is based on:

- Andrew B. Kahng, Seokhyeong Kang, **Seungwon Kim**, Kambiz Samadi and Bangqi Xu, “Power Delivery Pathfinding for Emerging Die-to-Wafer Integration Technology”, *Proc. IEEE/ACM Design, Automation and Test, in Europe*, 2019, pp. 842-847.

- Chapter V is based on:

- **Seungwon Kim**, Ki Jin Han, Youngmin Kim and Seokhyeong Kang, “Power Integrity Coanalysis Methodology for Multi-Domain High-Speed Memory Systems” *IEEE Access* (2019) to appear.
- **Seungwon Kim**, Ki Jin Han, Youngmin Kim and Seokhyeong Kang, “Fast Chip-Package-PCB Coanalysis Methodology for Power Integrity of Multi-domain High-Speed Memory: A Case Study”, *Proc. IEEE/ACM Design, Automation and Test, in Europe*, 2018, pp. 885-888.
- Byoungjin Bae, **Seungwon Kim**, Youngmin Kim, Seokhyeong Kang, Il Joon Kim, Kwangseok Kim, Sunwon Kang and Ki Jin Han, “A Preliminary Analysis of Domain Coupling in Package Power Distribution Network”, *Proc. IEEE International Symposium on Radio-Frequency Integration Technology*, 2017, pp. 19-21.

My coauthors have all kindly approved the inclusion of the aforementioned publications in my dissertation.

VITA

1990	Born, Seoul, South Korea
2014	B.Sc., Electrical Engineering, Ulsan National Institute of Science and Technology, Ulsan, South Korea
2019	Ph.D., Electrical Engineering, Ulsan National Institute of Science and Technology, Ulsan, South Korea

- **Seungwon Kim**, Ki Jin Han, Youngmin Kim and Seokhyeong Kang, “Power Integrity Coanalysis Methodology for Multi-Domain High-Speed Memory Systems” *IEEE Access* (2019) to appear.
- Andrew B. Kahng, Seokhyeong Kang, **Seungwon Kim**, Kambiz Samadi and Bangqi Xu, “Power Delivery Pathfinding for Emerging Die-to-Wafer Integration Technology”, *Proc. IEEE/ACM Design, Automation and Test, in Europe*, 2019, pp. 842-847.

*The papers have authors listed in alphabetical order.

- **Seungwon Kim**, Ki Jin Han, Youngmin Kim and Seokhyeong Kang, “Fast Chip-Package-PCB Coanalysis Methodology for Power Integrity of Multi-domain High-Speed Memory: A Case Study”, *Proc. IEEE/ACM Design, Automation and Test, in Europe*, 2018, pp. 885-888.
- **Seungwon Kim** and Youngmin Kim, “Analysis and Reduction of the Voltage Noise of Multi-layer 3D IC with Multi-paired Power Delivery Network”, *IEICE Electronics Express* 14(18) (2017), pp. 1-9.
- Mingyu Woo, **Seungwon Kim** and Seokhyeong Kang, “GRASP based Metaheuristics for Layout Pattern Classification”, *Proc. IEEE/ACM International Conference on Computer-Aided Design*, 2017, pp. 512–518.
- Byoungjin Bae, **Seungwon Kim**, Youngmin Kim, Seokhyeong Kang, Il Joon Kim, Kwangseok Kim, Sunwon Kang and Ki Jin Han, “A Preliminary Analysis of Domain Coupling in Package Power Distribution Network”, *Proc. IEEE International Symposium on Radio-Frequency Integration Technology*, 2017, pp. 19-21.
- **Seungwon Kim**, SangGi Do and Seokhyeong Kang, “Fast Predictive Useful Skew Methodology for Timing-Driven Placement Optimization”, *Proc. ACM/IEEE Design Automation Conference*, 2017, pp. 55:1–55:6.

- SangGi Do, **Seungwon Kim**, and Seokhyeong Kang, “Skew Control Methodology for Useful-skew Implementation”, Proc. International SoC Design Conference, 2016, pp. 221–222.
- **Seungwon Kim**, Youngmin Kim and Ki Jin Han, “Identification of Parameter Domain for the Design of High-speed I/O Interface”, Proc. IEEE Electrical Design of Advanced Packaging and Systems, 2015, pp.67–70.
- Jaemin Lee, **Seungwon Kim**, Youngmin Kim and Seokhyeong Kang, “An Optimal Operating Point by using Error Monitoring Circuits with An Error-Resilient Technique”, Proc. IFIP/IEEE International Conference on Very Large Scale Integration, 2015, pp.69–73.
- **Seungwon Kim**, Seokhyeong Kang, Ki Jin Han and Youngmin Kim, “Novel Adaptive Power Gating Strategy and Tapered TSV Structure in Multi-layer 3D IC”, *ACM Transactions on Design Automation of Electronic Systems* 21(3) (2016), 44.
- **Seungwon Kim**, Seokhyeong Kang, Ki Jin Han and Youngmin Kim, “Novel Adaptive Power Gating Strategy of TSV-based Multi-layer 3D IC”, *Proc. IEEE International Symposium on Quality Electronic Design*, 2015, pp. 537-541.
- **Seungwon Kim**, Ki Jin Han, Seokhyeong Kang and Youngmin Kim, “Analysis and Reduction of Voltage Noise of Multi-layer 3D IC with PEEC-based PDN and Frequency-dependent TSV Models”, *Proc. IEEE International SoC Design Conference*, 2014, pp. 124-125.

Chapter I

Introduction

In the recent trends of electronic applications including internet of things (IoT) and wearable devices, a main design concern is power management for lowering power consumption with maintaining the required power integrity. Although the low-power design can be realized in front-end-of-line (FEOL) level by reduced power supply complementary metal–oxide–semiconductor (CMOS) technologies [96], the overall low-power system performance is available with a proper design of power delivery network (PDN) for chip modules.

1.1 Difficulties in Power Delivery in Modern VLSI Design

Modern very-large-scale integration (VLSI) technology has gained higher performance and efficient power management based on advanced transistor technology. However, as shown in Figure 1.1, the cost reduction estimation beyond 20 *nm* node no longer be realized with advancing technologies. Only scaling chips by advanced technology nodes in FEOL no longer guarantee directly proportional benefits improvement. The reason why the relative cost reduction is reversed is that the yield of modern VLSI chips is very low, as advanced technology must satisfy complicate design rules during physical design process in back-end-of-line (BEOL). Moreover, after below 10 *nm* node, there is a big gap between resistance/capacitance delay of interconnection and delay of intrinsic transistor as shown in Figure 1.2. It is because advanced technology use a narrow metal stripe to connect standard cell to the standard cell. Also, high pin density of each cell makes difficulties on routing of interconnection. Thus, BEOL physical design is emerging to overcome the Moore's law.

In the industrial field, multi-die multi-chip-module (MCM) integration has been proposed to solve the yield and cost issue for high-performance CPU [3, 4, 5]. The traditional MCM breaks down monolithic

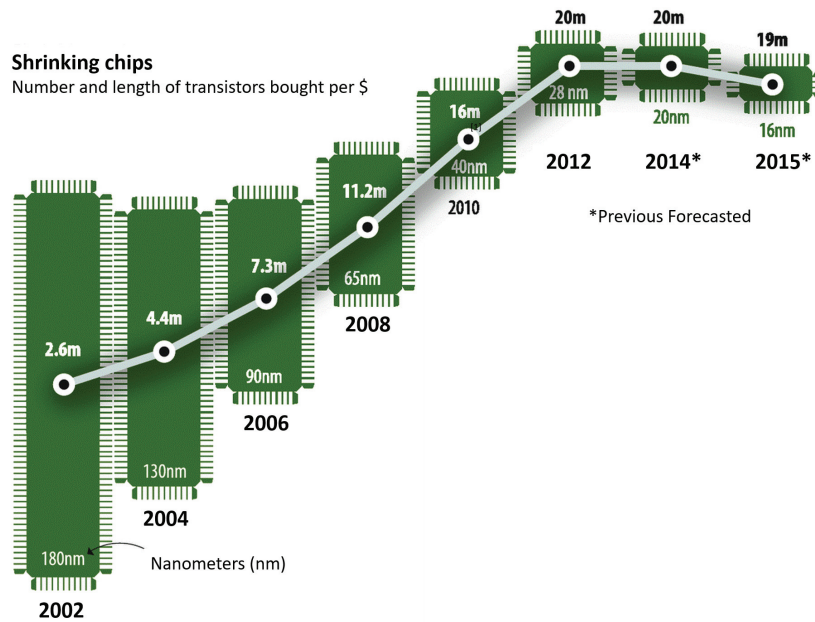


Figure 1.1: Changes in the integrated number of transistors per unit cost with the technology nodes in the VLSI chip. The costs beyond 20 nm node are predicted values [93].

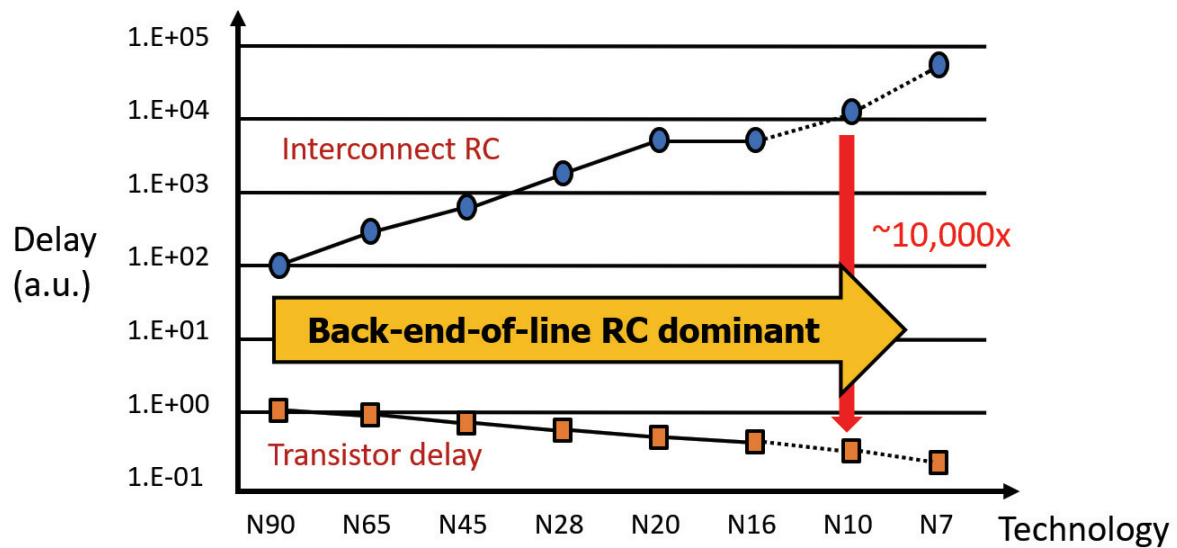


Figure 1.2: Growing gap between transistor delay and interconnect delay in advanced technology nodes [2].

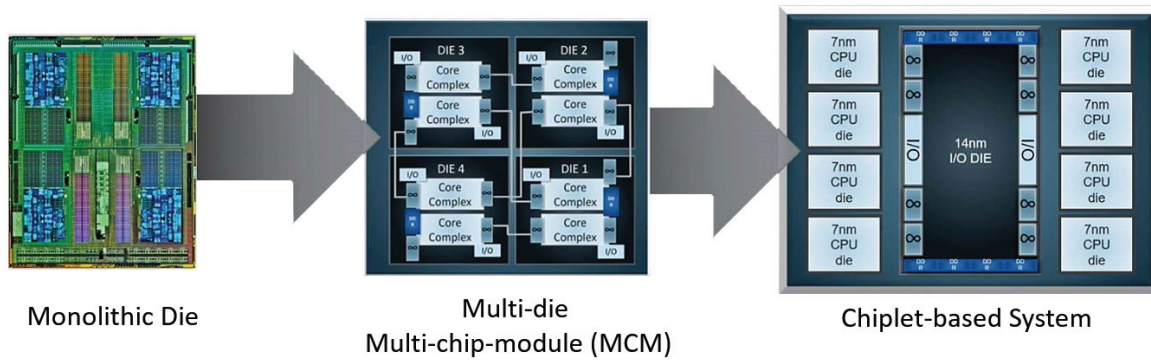


Figure 1.3: Multichip architecture revolution from monolithic die to *chiplet*-based system [6]

processors into separated chips. However, MCM approach still has limitations of bandwidth and latency related performance. Moreover, large power I/O and analog circuitry make system-on-a-chip (SoC) difficult to produce in advanced technology. Thus, modularization of function of die has been applied to solve the yield and cost problem in industrial field. The advanced modularized structure from MCM is called *chiplet* as shown in Figure 1.3. The *chiplet*-based systems separate *chiplet* functions, and each *chiplet* is optimized for their own functionality and process technology to minimize the *chiplet* size, improve yield, and reduce costs. In addition, *chiplet* also can reduce costs further by recycling known-good-dies [7, 8, 60, 61].

In a modern VLSI designs, heterogeneous architectural structures (see Figure 1.4(a)) and various 3D integration methods (see Figure 1.4(b)) have been used in a hybrid manner. Recently, the industry has combined 3D VLSI technology with the heterogeneous *chiplet*-based system. The 3D heterogeneous architectural structure is growing attention because it reduces costs and time-to-market by increasing manufacturing yield with high integration rate and modularization. However, power management is a major concern for lowering power consumption with maintaining the required power integrity from IR drop in 3D heterogeneous integration. As shown in Figure 1.4(b), some types of up-to-date high-performance SoCs use multiple 3D integration methods at once. For example, a 3D SoC with two dies combined should be co-designed for power planning with consideration of the IR drop due to vertical power delivery.

As the supply voltage has decreased below 1.0 V, the current demands for devices have increased. The supply noise caused by the product of current passing through IR drop in PDN of BEOL has become a critical problem for robust circuit operation. The demand for high-speed and high-performance integrated circuit (IC) operation with a low supply voltage and high IC switching current has resulted in a significant amount of supply-voltage fluctuations in the PDN [1, 10, 11, 12, 96]. Thus, PDN is an

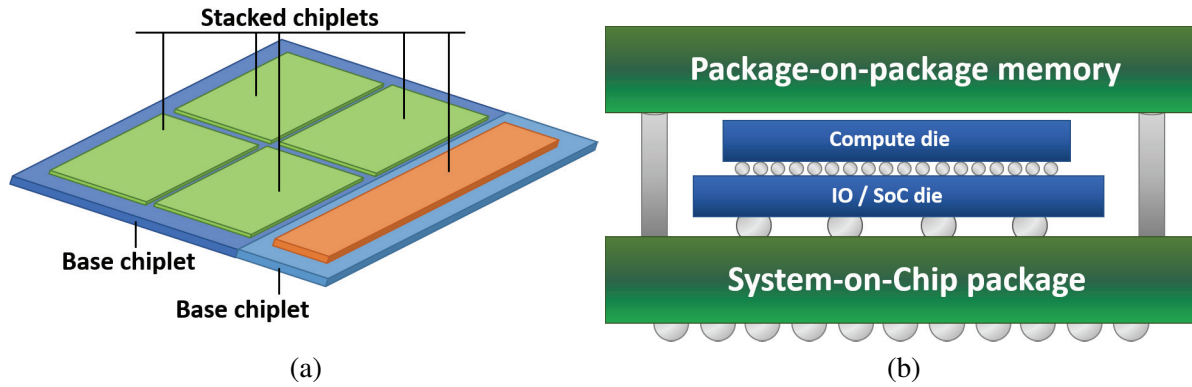


Figure 1.4: (a) An example of 3D *chiplet* architectural structure [94], and (b) side view of hybrid heterogeneous integration of 3D system-in-package and 3D system-on-chip [95].

integral aspect of physical design that directly affects reliability and functionality of product designs. A human designer can find ‘good’ PDN manually with iterative search, however, we do not know how much of this ‘good’ PDN is optimized in high design solution space. Moreover, the PDN design still depends on the engineer’s know-how. Thus, with increasing power density and complexities in modern VLSI designs, determining a high-quality PDN at the *Pareto frontier* of tradeoff between IR drop and chip routability is challenging (see Figure 1.5).

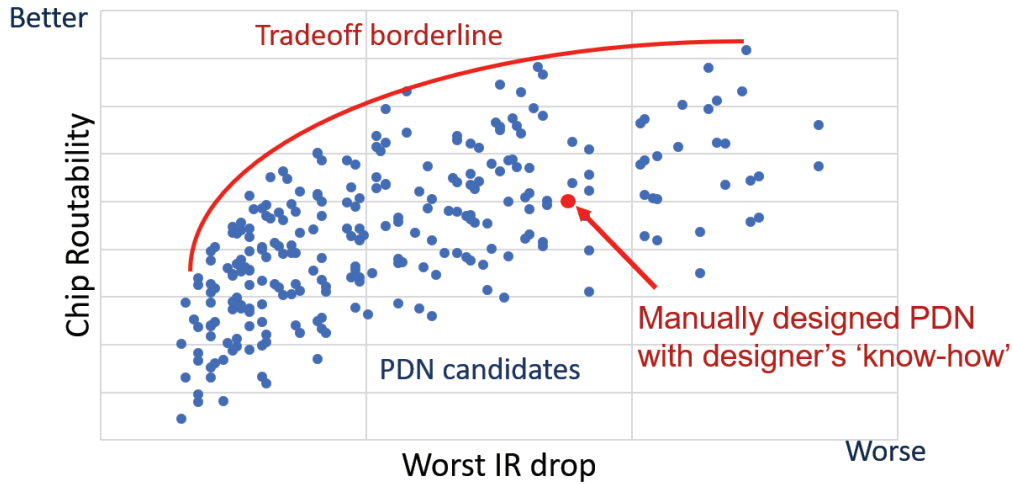
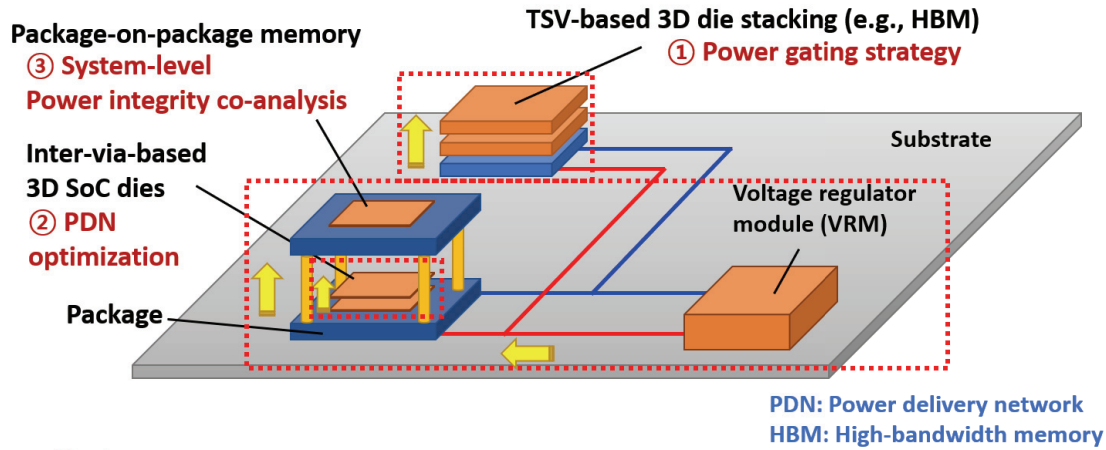


Figure 1.5: The on-chip routability versus IR drop envelope in high design solution space of on-chip PDN.

1.2 This Dissertation

In this dissertation, the innovative coanalysis and optimization methodologies have been proposed to solve power delivery issues for various types of integration methods applied to modern VLSI. First, power gating control is becoming more complicated as more dies are stacked in TSV-based 3D IC. Therefore, a new power gating control method considering IR drop dependency due to vertical 3D stacking is required. Second, in die-to-wafer (D2W) 3D IC implementation, the PDN is crucial to meet the design specifications. However, determining the optimal PDN design is nontrivial. On the one hand, to meet the IR drop requirement, denser power mesh is desired. On the other hand, to meet the timing requirement for a high-utilization design, more routing resources should be available for signal routing. Additional competition between signal routing and power routing is caused by inter-tier vertical interconnects in 3D IC. Thus, there is a need for a technique to find an optimal PDN in a one-pass flow without iteration effort of human engineers. Third, by using case-specific design models in a high-speed system integration, only a limited analysis of the effects of parametric variations can be performed in complex design problems, such as adjacent voltage domain coupling at high frequencies. Moreover, a conventional industrial method can be simulated only after completing the design layout; therefore, a number of iterative back-annotation processes are required for signoff; this delays the time to market. Therefore, system-level power integrity analysis technology combining high-flexibility model is required. In each of these three thrusts, this dissertation proposes novel analysis and optimization methodologies for power delivery in various modern VLSI integrations. Figure 1.6 illustrates the scope



- ① Chapter 3: Power gating strategy
→ Minimize IR drop and wake-up time
- ② Chapter 4: PDN optimization
→ Identify optimal PDN configurations that offer best routability
- ③ Chapter 5: System-level power integrity co-analysis
→ System-level Fast analysis methodology with considering multiple VDD domain

Figure 1.6: Scope of this dissertation.

of this dissertation.

The remainder of this dissertation is organized as follows.

- Chapter II covers background and reviews prior works in the area of power delivery analysis and optimization in various modern VLSI integrations. We explore the background for 3D integrations of modern VLSI designs and previous work to optimize IR drop due to vertical power delivery in 3D IC.
- Chapter III covers our investigation of the voltage noise in a multi-layer 3D IC stacking with PEEC-based on-chip PDN and frequency-dependent TSV models. We propose a wire-added multi-paired on-chip PDN structure to reduce voltage noise in a 3D IC. In addition, we propose a novel power gating strategy that optimizes the in-rush current profile, subject to the voltage-drop constraints. Moreover, a tapered TSV architecture based on the layer dependency has been analyzed.
- Chapter IV proposes a power delivery pathfinding methodology for emerging die-to-wafer integration, which seeks to identify an optimal or near-optimal PDN for a given design and PDN specification. Our pathfinding methodology exploits models for routability and worst IR drop,

which helps reducing iterations between PDN design and circuit design in 3D IC implementation. We present validations with real design examples and a 28 nm foundry technology.

- Chapter V covers the preliminary analysis of our multiple power domain PDN model. In addition, we propose a power integrity coanalysis methodology for multiple power domains in high-frequency memory systems. Our proposed methodology can analyze the tendencies in power integrity by using parametric methods such as parameter sweeping and Monte Carlo simulations. We also prove that our proposed methodology can predict similar peak-to-peak ripple voltages that are comparable with the realistic simulations of low-power double data rate four interfaces.
- Chapter VI summarizes key contributions of this dissertation and presents future directions for optimization of power delivery in future VLSI designs.

Chapter II

Background and Prior Work of Power Delivery Optimization in Modern VLSI Designs

This chapter covers background and reviews prior works in the area of power delivery analysis and optimization in various modern VLSI integrations. We explore the background for 3D integrations of modern VLSI designs and previous work to optimize IR drop due to vertical power delivery in 3D IC.

To ensure power integrity in modern VLSI designs, semiconductor designers have been analyzing and optimizing BEOL's power delivery at on-chip level first. 3D IC stacking technologies have emerged as the main hope for enhance system integration, reduce the area footprint, and extend design performance/power envelope for modern VLSI designs [13, 14, 15, 96]. The 3D IC effectively reduces the propagation delay by shortening the wire lengths using vertical interconnections. This section introduces background and prior works on power delivery optimization techniques applied to various integration methods proposed in this dissertation.

2.1 Power Gating

Power gating is a technique that drastically reduces the leakage power by cutting off the current path between the supply and the ground with (high- V_{th}) switching transistors [16, 17]. During the idle mode, the switching transistors turn off, and the internal logic power (virtual power) is disconnected from the main supply power (see Figure 2.1). Whenever circuit operation needs to be resumed, enable signals

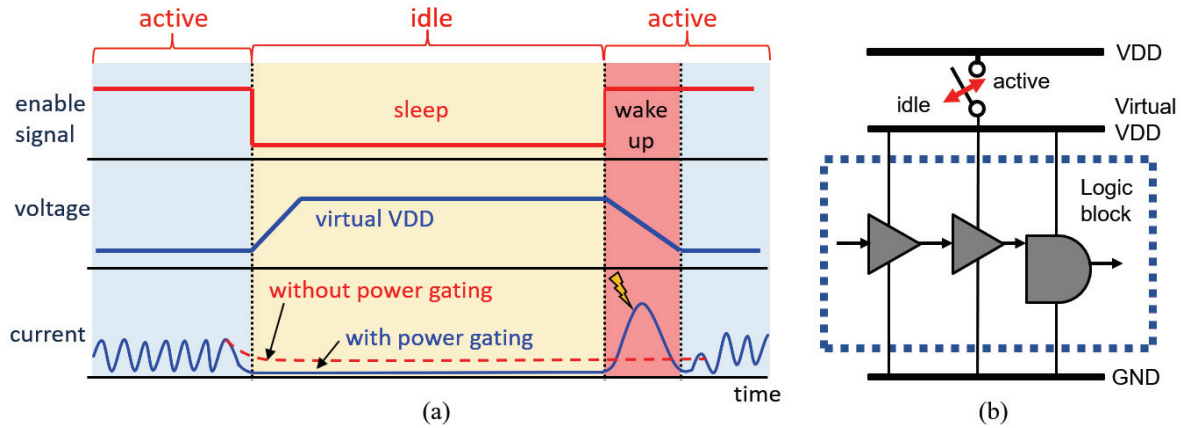


Figure 2.1: (a) Power gating operation with (b) PMOS header switch [79].

turn on the switches to wake up the internal logic gates.

Power Gating Potential Hu et al. [18] introduces microarchitectural techniques for power gating. They firstly develop parameterized analytical equations for estimating the break-even point for application of power gating; the power gating potential equation results in floating-point and fixed-point units being put to sleep for up to 28% and 40% of the execution cycles, respectively.

Architectural-driven To let a processor cut off power frequently, previous research [19, 20, 21, 22] uses Multi-Threshold CMOS as the power gating footer switches to support multiple sleep modes which are categorized as the tradeoff between wake-up overhead and leakage saving. A power gating technique supporting multiple sleep modes with a robust gate-bias generator is addressed in [19]; the multiple modes provides an extra 17% reduction in overall leakage as compared to single mode. Singh et al. [20] analyzes the impact of datapath partitioning and proposes a fine-grained soft gating method resulting in greater leakage power savings. To handle major components of leakage current both in logic and memory circuits while ensuring minimum ground bouncing, a power gating including an additional power gating path in parallel to conventional sleep transistors is introduced in [21]. Zhang et al. [22] makes improvements on the analog power switching structure to simple structure without analog components; this proposed structure shows high tolerance to manufacturing process variation and low static power.

Scheduling Approach During wake-up operation, however, a very large amount of current (i.e., an in-rush current) can flow through the power supply lines when many switches are simultaneously turned

on in the chip as shown in Figure 2.1(a). As a result, circuits that are already operating will experience a significant voltage fluctuation because of the IR drop in the metal interconnects; this reduced voltage affects the circuit performance. A power gating aware task scheduling in multiprocessor system-on-chip (MPSoC) has been suggested by [24]. The authors achieve 25% performance improvement with up to 80% noise protection. Another research suggests a wake-up scheduling method considering both resource usage and timing budget for power gating [25]. The authors propose an efficient algorithm to find a wake-up scheduling, which achieves a significant hardware resource reduction with little impact on the wake-up time. To catch up the best tradeoff between wake-up and inrush current, Jeong et al. [79] proposes programmable power gating switch (PPGS) design which can divide wake-up cells into two groups (*few* and *rest*); the first stage *few* turns on to allow the limit charge current, and then the remaining *rest* turn on. Kahng et al. [23] claims that [79] does not consider out-of-order execution, the benefits of exploiting core location and stage information to determine safe wake-up modes; they propose token-based adaptive power gating technique which considers these issues and shows the importance of stagger to reduce core wake-up latency.

Power gating can also be applied to 3D ICs, however, there is no study that considers the IR drop differential due to vertical stacking. In Chapter III, we investigate the in-rush current in the power gating of a 3D IC coupled with frequency-dependent tapered TSV models and propose an adaptive power gating technique to maximize performance by reducing wake-up time.

2.2 Power Delivery Network in On-Chip

As the size and the length of the inter-tier vias (VIs) become very smaller, the density of the circuit increases but the IR drop issue has emerged due to higher resistance of vertical interconnection more than TSVs. Moreover, PDN competes with signal routing VIs for a restricted budget of resources and cost, these competitions may cause a breakdown of design implementation and performance degradation.

Packaging-driven 3D IC PDN Regarding *packaging-driven* 3D IC PDNs, Healy et al. [11] investigated various properties of the power system architecture in the 3D IC. They reveals a limitation of the dynamic noise for a given turn-on time. Further, they used simplified Resistance, Inductance, Conductance and Capacitance (RLGC) models for the TSV and ignored the on-chip interconnections. Huang et al. [10] proposed a physical model for the PDN with 3D IC stacking. The voltage drops in the different layers were analyzed using a decoupling capacitor. Jung et al. [12] analyzed the impact of power and

ground (P/G) TSVs in 2D and 3D full-chip layouts. Various P/G TSV placements and patterns were simulated to determine the optimum P/G TSV locations. Xu et al. [26] proposed a compact RLGC model for the TSV in a 3D IC. They showed that the inductance component of the TSV has the largest impact on the VDD/GND noise. Kim et al. [27] proposed a scalable electrical model of a TSV, including all possible parasitic effects of the TSV-last process, and carried out a time-domain analysis using an eye diagram. Khan et al. [28] investigated the IR drop noise in a 3D IC. The resistance of the TSV and on-chip PDN resulted in a significant impact on the voltage drop in a 3D IC. He et al. [29] proposed compact models of the IR drop in a 3D IC PDN. They claimed that the voltage drop has a quadratic dependency on the number of chips in a stack.

Foundry-driven 3D IC PDN Conventional *packaging-driven* 3D IC integration technologies with TSVs are limited by TSV size and pitch, which constrains achievable vertical integration density [38]. Therefore, Multiple *foundry-driven* 3D integration technologies have recently emerged as viable solutions with significant PPAC benefits; these include high-precision face-to-face (F2F) wafer-on-wafer (WoW) and die-to-wafer (D2W) stacking [43, 56] (see Figure 2.2). WoW technology is more tailored towards power / performance / area improvement, whereas D2W aims to provide more cost-effective integration while also providing system-level power / performance improvements (e.g., for memory-on-logic, single-chip solutions, etc). WoW faces two key limitations compared to D2W technology: (1) same area constraint for top and bottom dies, which limits partitioning scenarios, and (2) lack of commercial EDA support. On the other hand, all IPs in the D2W technology are still 2D and are only partitioned across multiple dies (e.g., a large bottom die and various-sized smaller top dies). Hence, there is no need for special 3D EDA support in the D2W regime. This flexibility, coupled with relatively high integration density, has made the D2W technology a practical solution to cope with 2D scaling challenges.

The challenges are exacerbated in *foundry-driven* 3D ICs because of additional resistance between the power source and transistors in different tiers. Further, as shown in the Figure 2.3, inter-tier vertical interconnects (VIs) must support both signal and power / ground routing, which limits feasible integration. At the same time, as VIs become smaller to support higher integration densities, they become more resistive, with adverse effects on PDN [58, 59]. To achieve robust functionality of 3D ICs, designers must mitigate and balance the aforementioned PDN-related challenges. This demands an efficient, accurate design space exploration (a.k.a. pathfinding) methodology that – given various technologies and design-dependent parameters – can quickly provide quality of result (QoR) tradeoffs of various PDN solutions.

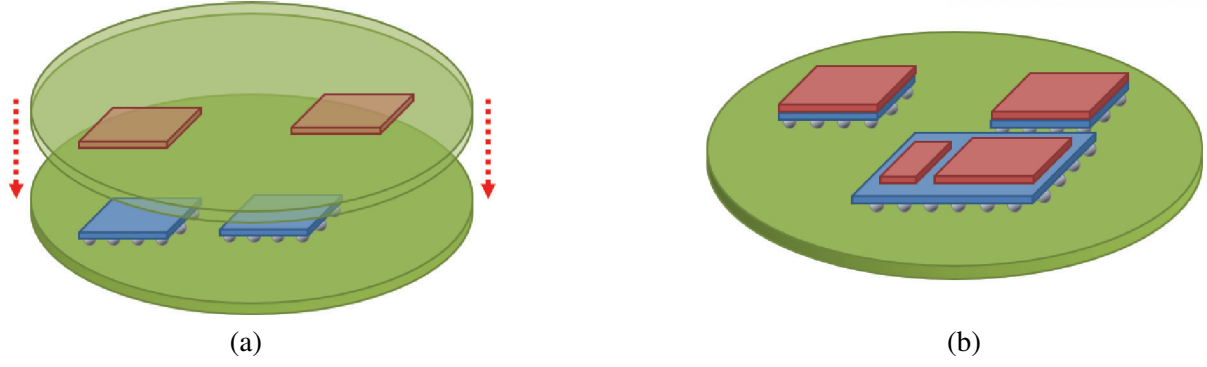


Figure 2.2: Examples of 3D IC integration technologies. (a) wafer-on-wafer integration, and (b) die-to-wafer integration.

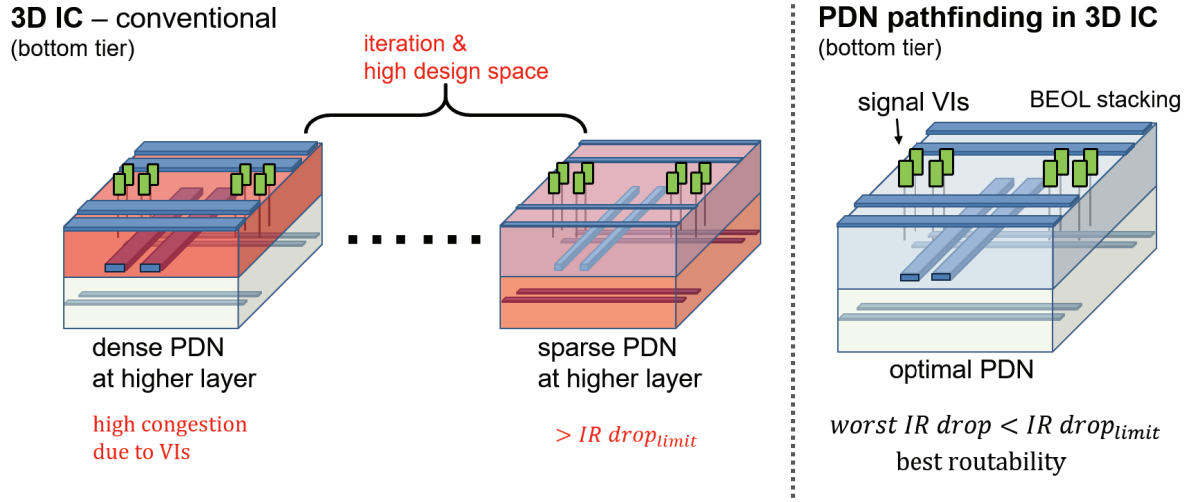


Figure 2.3: An example of a PDN design of one die in a 3D IC containing VIs.

In Chapter IV, we propose a machine learning based PDN optimization methodology for emerging D2W integration to identify a nearly optimal PDN for a given design and PDN specification.

2.3 System-level Power Delivery Analysis

Recently, the data rate and clock speed of high-speed I/O interfaces, such as mobile memory, have increased beyond the gigabit-per-second level. However, although target performances have been met, certain undesirable effects, such as mismatch and crosstalk, have also occurred in systems. In addition, physical limitations reveal low power efficiency with increasing of input / output (I / O) bandwidth between the central processing unit and memory [67]. Therefore, we need to maximize the performance of the electrical links in a printed circuit board (PCB) and the package structure in high-speed signal

transmission conditions.

The packaging structure causes static and dynamic power losses. In addition, the memory system requires additional on-chip power consumption to ensure signal integrity, such as on-die termination and an equalizer. Therefore, in a high-speed memory system that requires a low-power mobile device, the package structure needs to minimize the power loss to ensure power integrity (PI). It is important to analyze the effect of the decoupling capacitors on the dynamic power loss because the PDN in the system also causes frequency-dependent IR drops [68]. In the state-of-the-art mobile and wearable devices, we need to consider coupling the separate power domains as well. The current low-power memory systems apply multiple power domains to deliver power with adequate supply voltage (VDD) levels. Each power domain has its own noise and switching activity [69]. For example, the high-speed low-power double data rate four (LPDDR4) memory has three power domains: VDD1 and VDD2 for the core and VDDQ for the I/O buffers [70]. However, previous multi-domain studies have focused on case-specific design analysis or the lack of coanalysis on the package in the system [73, 72, 71].

A prediction of power integrity in high-frequency systems is difficult; therefore, to overcome the signoff constraint and improve results, the industry has designed chip-package-PCBs by numerous back-annotations instead of efficient codesigns. It is time-consuming to simulate the full layout of the memory system using a realistic package and PCB model with EDA tools; a complete layout is also required [71, 74]. Therefore, to reduce the number of iterations in the design process, it is necessary to analyze the effect of the design variables that constitute the memory system on the PI. However, in previous case-specific design models [73, 72], it was difficult to construct a new model when the memory system variables changed. In the simplified models at the system-level [71], it was difficult to model the parameters required for a practical package design for the memory system. Therefore, if the practical extracted parameters and the numerical model are compatible with each other in one analysis methodology, it is possible to have fast and massive parametric simulations at the proper level of accuracy.

In Chapter V, we propose a PI-aware coanalysis methodology with consideration of the effects of the electrical and structural parameters on the multiple power domain chip-package-PCB system.

Chapter III

Adaptive Power Gating Strategy and Tapered TSV in Multi-Layer 3D IC

In this chapter, we first propose a simple and accurate analysis methodology to investigate the PDN of 3D IC stacking structures. The parasitic of on-chip metal PDN is extracted with partial electrical equivalent circuit (PEEC)-based models [30]; and frequency-dependent TSV parameters, including C4 bumps, are generated by the electromagnetic (EM) modeling method [31, 32, 33]. We combine the on-chip PDN and TSV in multi-layer 3D IC for power integrity analysis of the static and dynamic voltage drop. In addition, various PDN structures are investigated in the proposed 3D IC analysis methodology, and an optimal PDN architecture for reducing IR drop is obtained. Figure 3.1 shows the general TSV-based (*packaging-driven*) 3D IC stacking methodology of two layers. The multiple dies are stacked in face-to-back bonding to stack more than two layers. We vary the number of stacking layers to identify the impact of the multi-layer on the power integrity.

We also investigate the power gating of TSV-based 3D IC stacking structures. We propose an adaptive two-stage power gating strategy in a 3D IC with interval finding algorithm and analyze the tapered TSV architecture to reduce the wake-up time while satisfying the IR drop constraint. The main contributions in this chapter are summarized as follows:

- We investigate various PDN structures in the proposed 3D IC analysis methodology, and propose an optimal PDN to reduce IR drop.
- We use 3D EM based modeling methodology to generate frequency-dependent scattering parameter (S -parameter) for TSV structures.

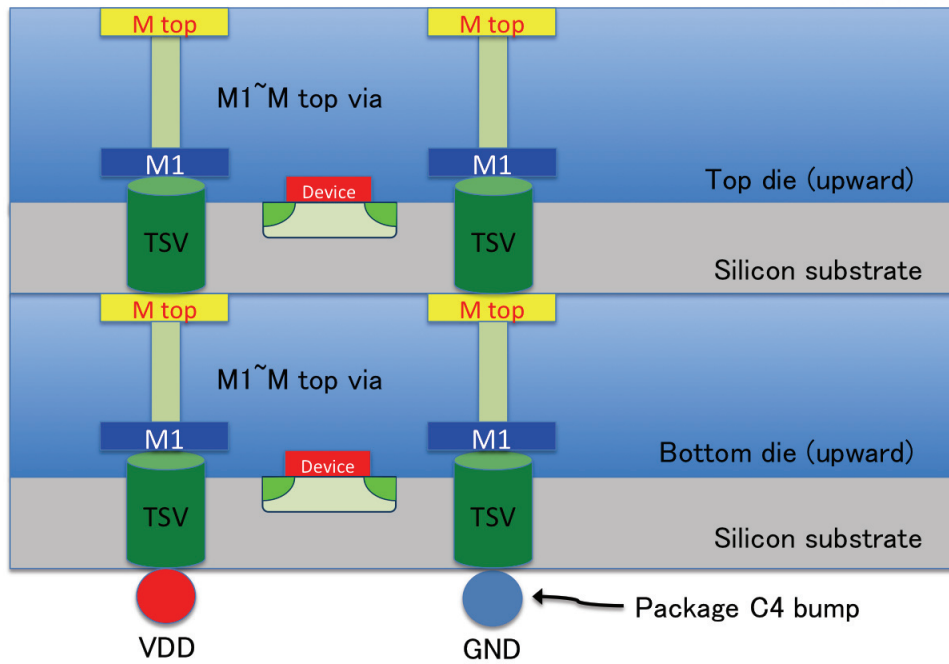


Figure 3.1: TSV-based face-to-back bonding of the multi-layer 3D IC.

- We combine the on-chip PDN and TSV in a multi-layered 3D IC to perform power gating analysis of the static and dynamic voltage drops and in-rush current.
- We investigate the layer-dependent IR drop problems in the power gating of the 3D IC using frequency-dependent TSV models.
- We propose a novel power gating strategy that optimizes the in-rush current profile in a 3D IC, subject to supply noise constraints. The proposed method can achieve a faster wake-up time and satisfy the IR drop requirements.
- A algorithm to find the optimum interval between switch-enable signals of the 3D IC based on the operation status of each layer is proposed.
- We have analyzed the tapered TSV architecture in our proposed method so as to further reduce the wake-up time and balance the intervals among layers.

3.1 Preliminary Analysis of Power Delivery Network

3.1.1 Analysis of power delivery network in the TSV-based 3D IC

As the PDN has become complicated, and wire resistance has increased owing to interconnect scaling, supply voltage fluctuations due to the IR drop have become a major issue in PDN design. A PEEC method [30] has been introduced for an accurate extraction of the RLC components of wires. We use the PEEC method to extract the RLC components from a PDN interconnect mesh having nine VDD and nine GND pads [31]. To obtain a frequency-dependent model of TSV structures, we use a 3D EM method [32]. Frequency-dependent S -parameters of the TSV structures are extracted from the 3D EM modeling method based on a mixed-potential integral equation [32, 33].

We apply the EM modeling method to the TSV structure shown in Figure 3.2, and generate S -parameter data with TSV parameters, as described in Table 3.1 [96]. We then perform a dynamic transient analysis with clock buffers to simulate the voltage fluctuations, and analyze the PDN architecture in 3D IC stacking.

Table 3.1 summarizes the metal interconnect parameters and TSV dimensions used in this study. The on-chip PDN consists of two top metal lines (i.e., M6 and M5), and the power supply of the logic block is connected through vias to supply the required current as shown in Figure 3.3; (a) and (b) show 3D visualizations of the PEEC elements in the regular on-chip PDN structure and the multi-paired structure, respectively. All the wire segments and vias are plotted, and ports for the VDD and GND are shown in red color. The corresponding VDD and GND ports and the logic block is represented in Figure 3.3(c). We assume that the $0.5\text{mm} \times 0.5\text{mm}$ die contains nine VDDs and nine GNDs, and all of these ports are connected to the corresponding S -parameter TSV ports to construct the entire PDN of 3D IC, as shown in Figure 3.1. We apply an ideal DC voltage source simultaneously to all the power and ground TSVs. The center VDD and GND ports are connected to the power supply pins of the logic gates (i.e., clock buffers) as shown in Figure 3.3(c). The logic consists of 40 clock buffers to mimic the current load in the simulations, and is implemented in 22-nm technology node [100]. The nominal VDD value is set to 0.9 V, and the clock buffer size is $Wn = 4\ \mu\text{m}$ and $Wp = 8\ \mu\text{m}$. We then measure the power supply fluctuation for various configurations of 3D IC when the internal logic makes transitions in *SPICE* transient analysis [97]. First, we stack a single layer of dies with TSV and C4 bump. Then, we vary the number of stacking layers, to identify the impact of the multi-layer on the IR drop.

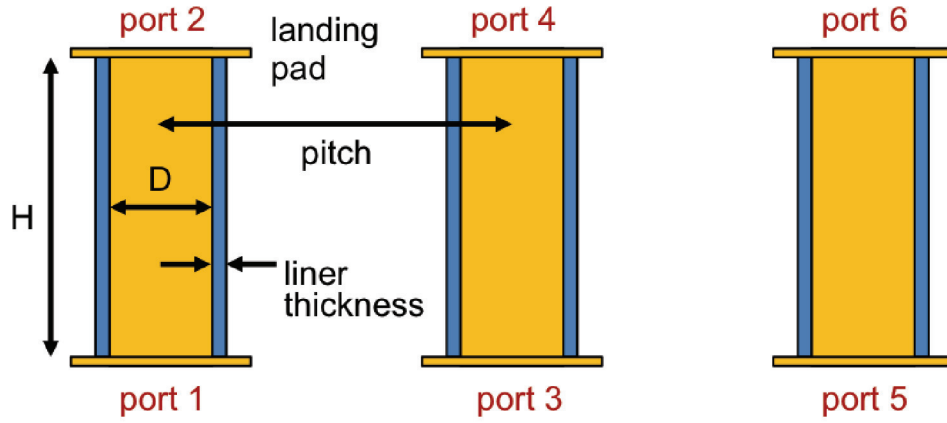


Figure 3.2: Power and ground TSV structure with six ports for frequency-dependent S -parameter generation [32].

Table 3.1: Structural parameters of the metal interconnect and TSV.

Parameter	Value (μm)
Metal 5 (M5) width	5
Metal 5 (M5) pitch	200
Metal 6 (M6) width	10
Metal 6 (M6) pitch	200
TSV diameter (D)	10
TSV height (H)	50
TSV pitch	200
TSV linear thickness	0.1
TSV pad size	15 x 15
Landing pad thickness	0.1
Silicon conductivity	10 S/m

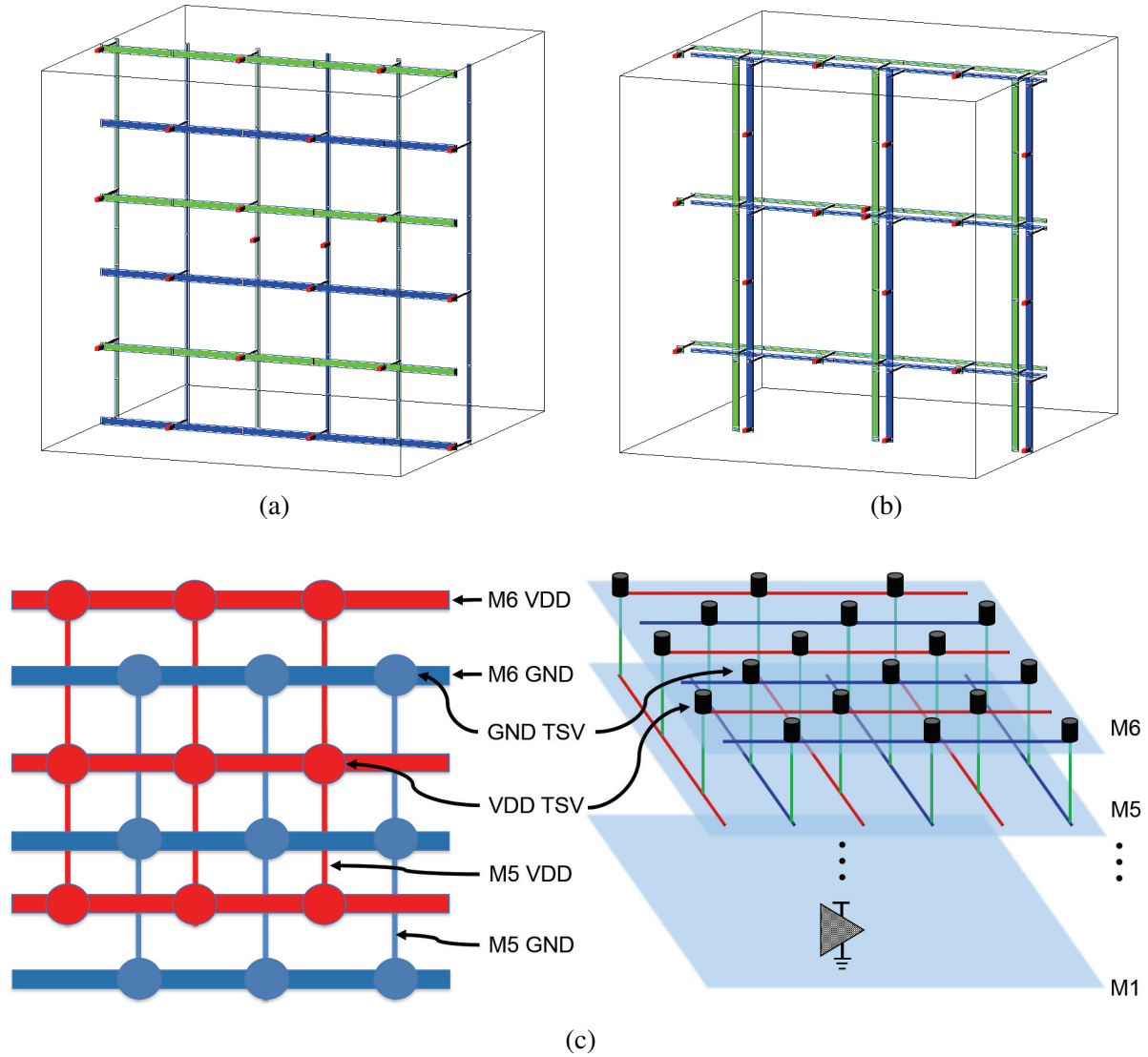


Figure 3.3: 3D view of (a) the regular and (b) the multi-paired PDN structure. (c) Top view and corresponding VDD and GND ports in M5 and M6. PDN is generated using M6 and M5 with 3×3 VDDs and GNDs. Green line is VDD and blue line is GND metal, respectively.

Table 3.2: The worst VDD droop in the conventional regular PDN and the proposed multi-paired PDN.

The number of layer	Regular structure VDD droop (mV)	Proposed structure VDD droop (mV)	Reduction (%)
1	88.8	70.2	20.9
2	122.3	109.1	10.8
3	146.0	136.0	6.8
4	164.4	155.7	5.3
5	178.5	170.8	4.3

3.1.2 Multi-paired PDN structure for reduction of voltage noise

Multi-paired PDN structure

Figure 3.4(a) shows the voltage fluctuation results of the 3D IC with a single layer TSV. As shown, the TSV attributes 89% additional voltage droop and 98% ground bump in a single layer. Figure 3.4(b) shows the impact of the number of 3D IC layers on the voltage drop in both RL and RLC PEEC models. As shown, larger voltage fluctuation is expected as the number of layer increases. In addition, the RLC model shows more impact on the voltage variation than the RC model. For example, approximately 1.8 times more VDD droop occurs for five layers 3D IC from single layer with all RLC included and the additional 10% (five layers) to 43% (single layer) droop attributes to large inductance in the PDN. Therefore, the inductance component should be considered in high-frequency 3D IC for accurate voltage fluctuation analysis, such as $L di/dt$ noise.

Shorter wire space between VDD and GND can effectively reduce the mutual inductance of wire [34]. Thus, we propose and investigate the multi-paired PDN structure shown in Figure 3.3(b), and compare the voltage fluctuation with that of the conventional regular structure (e.g., Figure 3.3(a)). We measure the benefit of the proposed PDN architecture with increasing number of layers in 3D IC stacking. Table 3.2 compares the VDD droop between the regular PDN and the proposed multi-paired PDN. As shown, the proposed PDN architecture achieves up to 21% reduction in VDD droop for a single layer, and the benefit is attenuated as the number of layers increases.

Wire-added multi-paired PDN structure

A narrower wire space between VDD and GND effectively reduces mutual inductance, and appears to provide less voltage fluctuation on PDN. Based on this result, we investigate the effect of the metal density and the wire space between wire pairs. First, additional VDD and GND wire pairs are added to the multi-paired PDN structure, as shown in Figure 3.5. Then we vary the wire space between the

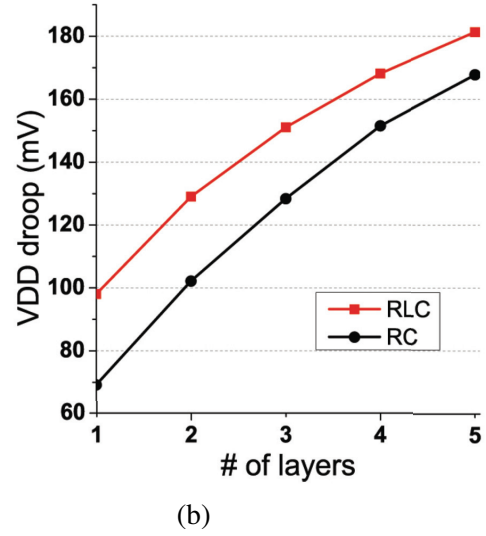
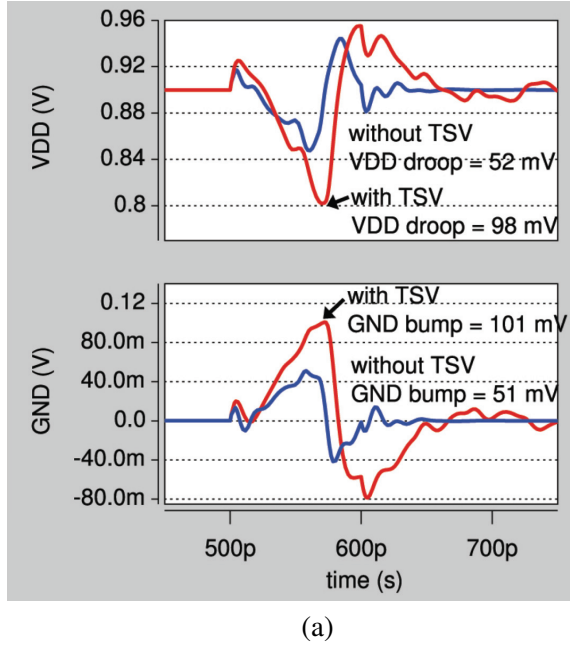


Figure 3.4: (a) VDD droop (top) and GND bump (bottom) comparison with TSV and without TSV in a single layer of die. (b) Impact of the number of layers and the inductance (L) on the PDN voltage noise.

original wires and newly added wire pairs. Figure 3.5(a) shows the minimum space between the original wire pairs and added wire pairs (i.e., $20 \mu m$). Then we increase this space to $140 \mu m$ in $20 \mu m$ steps, which is the minimum wire space of M6. The interconnect in M6 layer has larger width than M5, thus the same wire space can generate more symmetric structures in our study. Added wire pairs overlap the pre-existing original wires when the space becomes larger than $140 \mu m$. Also, $80 \mu m$ wire space is impossible, because the GND wire passes through the VDD TSV pad.

Table 3.3 summarizes the simulation results of the worst VDD droop for various wire space in the wire-added multi-paired structure. The reduction of the voltage droop is calculated with respect to the conventional regular structure (see Figure 3.3(a)). As shown, the case of $20 \mu m$ wire space (see Figure 3.5(a)), which is the minimum, results in the best reduction of the VDD droop, because it has the shortest return paths. On the other hand, the $120 \mu m$ wire space case, shown in Figure 3.5(b), provides the worst voltage fluctuation (and minimum reduction from the regular structure), because it has the longest current return paths. It is worth mentioning that the PDN structure that has more regular space between wires degrades the efficiency of inductance reduction (e.g., Figure 3.5(b)).

In addition, to analyze the benefit of the proposed wire-added multi-paired PDN on the power integrity of the 3D IC, we apply the optimum wire-added multi-paired structure (i.e., minimum wire space) to multi-layer 3D IC. Figure 3.6 shows that for symmetry, one more VDD and GND wire pair is added

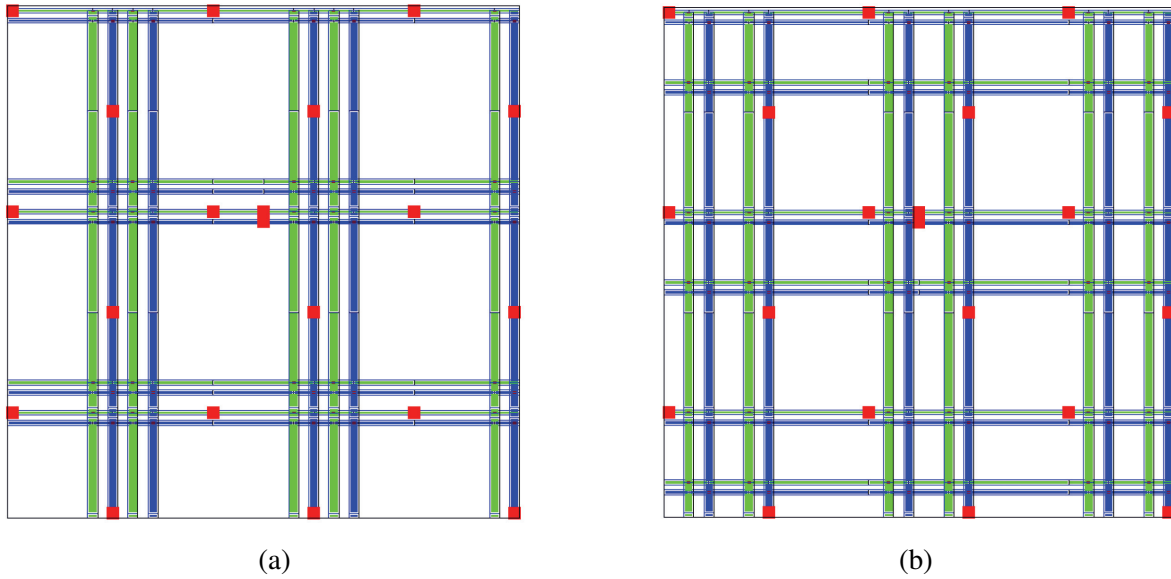


Figure 3.5: Top view of the wire-added multi-paired (two pairs) PDN structure of (a) $20\ \mu\text{m}$ (minimum) wire space, and (b) $120\ \mu\text{m}$ wire pair space.

Table 3.3: The worst VDD droop in the wire-added multi-paired (two pairs) PDN structure. $80\ \mu\text{m}$ space is impossible, because the GND line runs over the VDD ports in M6 layer.

Wire space (μm)	Worst VDD droop (mV)	Reduction (%) (ref. to regular structure VDD droop)
0 (no added wire)	70.2	20.9
20	64.8	27.0
40	65.7	26.0
60	66.9	24.7
100	69.4	21.8
120	71.2	19.8
140	67.0	24.5

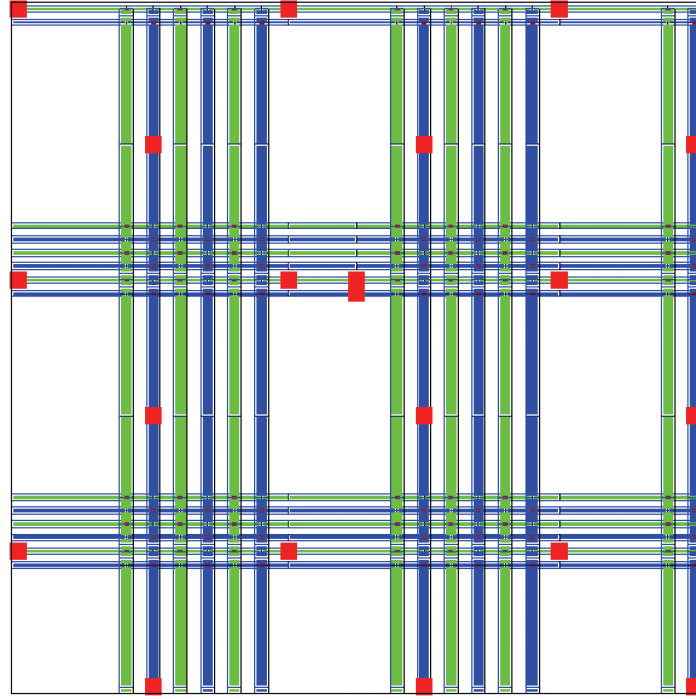


Figure 3.6: Top view of two VDD (green) and GND (blue) wires are added to the proposed multi-paired (three pairs) PDN structure.

to the final PDN structure. Figure 3.7 compares various PDN architectures. For a single layer, three pairs (two additional pairs) of VDD and GND reduce the VDD droop by more than 29%, compared to the regular (i.e., non-paired) PDN structure. However, the VDD droop reduction attenuates as the number of layer increases, showing approximately 29.1% to 6.8% reduction from one to five stacking layers. The benefit of the two wire-pair addition (three pairs) is only less than 2% points from the single wire-pair case. Therefore, a multi-paired (one pair) structure without adding any additional wire pairs is the best option to minimize voltage noise with limited wire resources. Two pairs of the multi-paired PDN structure is considered to be the best tradeoff between voltage noise reduction and wire resources.

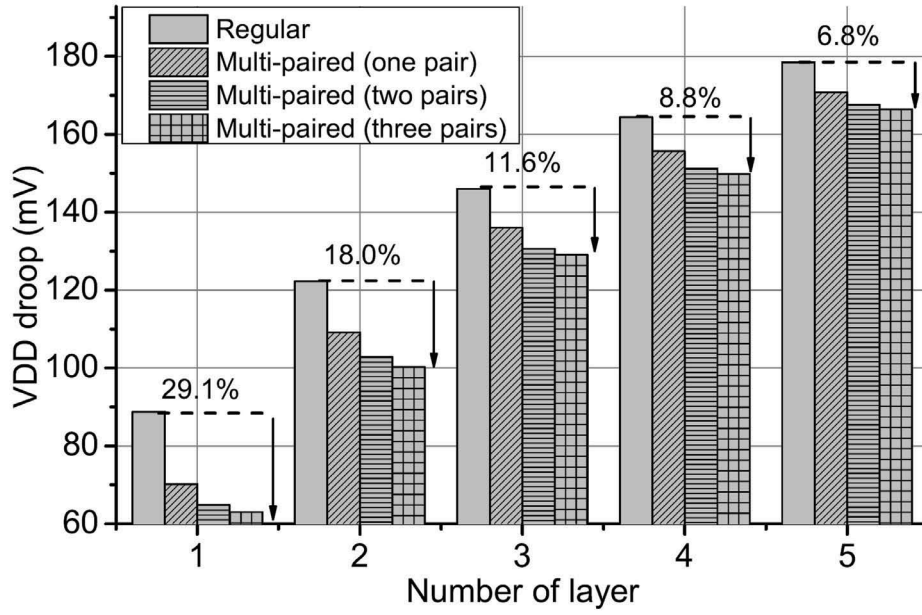


Figure 3.7: The worst VDD droop comparison in the conventional regular (non-paired) PDN and our various proposed multi-paired PDN structure.

3.2 TSV-Based 3D IC Power Delivery Network

3.2.1 TSV EM modeling and S -parameter extraction

To obtain a frequency-dependent model of TSV structures, we use a 3D EM method [32]. The frequency-dependent S -parameters of the TSV structures are extracted using a 3D EM modeling method based on a mixed-potential integral equation [32]. Because spatial discretization is avoided by employing the modal basis functions, the method produces a reduced system matrix and a simplified equivalent circuit [33] compared to the conventional PEEC method [30]. The circuit model obtained from the EM method can be converted into a set of multiple port network parameters. In this work, we apply the EM modeling method to the eighteen-TSV (nine VDD and nine GND) structure shown in Figure 3.2 and Figure 3.3. For the modeling, we generate thirty-six-port S -parameter data with the TSV parameters, as summarized in Table 3.1. The frequency-dependent insertion loss and return and coupling loss of the S -parameter transmission coefficients of the TSVs are plotted in Figure 3.8. All responses in Figure 3.8 follow the typical characteristics of TSV interconnects, in which low-frequency and high-frequency behaviors exhibit the slow-wave mode and the quasi-TEM mode, respectively.

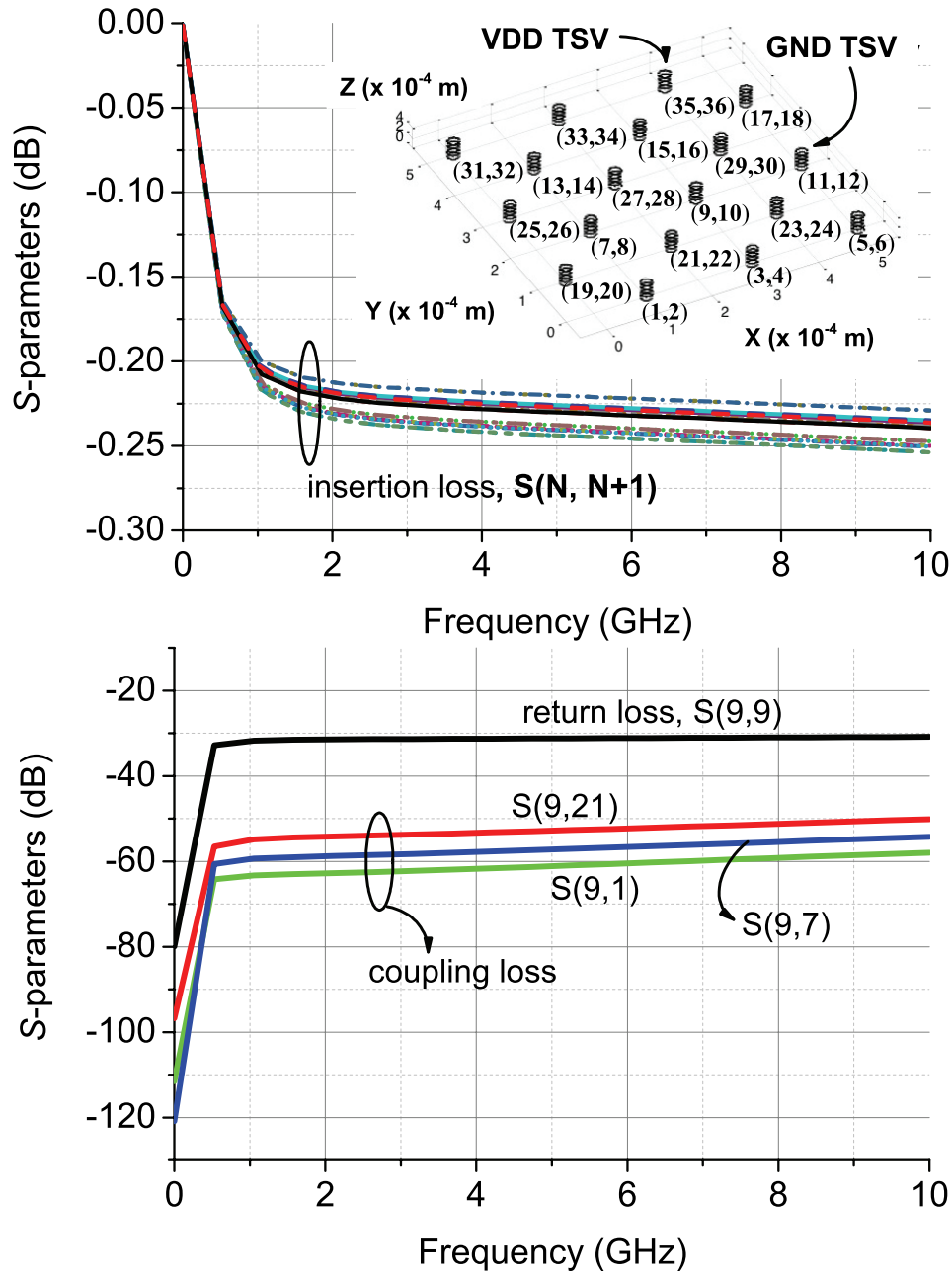


Figure 3.8: Insertion loss (top) and return and coupling loss (bottom) in S -parameter of the power TSVs as the frequency increases. The geometry of the 18 TSVs (9 VDD and 9 GND) is shown in the top figure, including port numbers.

3.2.2 PEEC-based on-chip PDN

As the PDN has become complicated and the wire resistance has increased because of interconnect scaling, the supply voltage fluctuations from the IR drop have become a significant problem in the PDN design. A PEEC method [30] has been introduced for accurate extraction of the RLC components of the wires. We use the PEEC method to extract the RLC components from a PDN with nine VDD and nine GND pads. We then perform a dynamic transient analysis with clock buffers to simulate the IR-drop-related voltage fluctuations and evaluate our power gating strategy in a 3D IC.

The metal interconnect parameters used in the PDN are summarized in Table 3.1. The PDN consists of two top metal lines (i.e., M5 and M6) and is connected to M1 through vias to supply the required current to the logic blocks. To extract the PEEC-based RLC components of the PDN structure, we implement a program with C++ and a standard template library (STL). Figure 3.3(b) shows a 3D visualization of the PEEC elements in the PDN structure. In the figure, all wire segments and vias are plotted; the ports for VDD and GND are shown in green. We use same PDN structures for all other layers. There are three layers of PDN metals (i.e., M6, M5, and M1) and vertical lines are vias to connect each metal layer. The blue lines are ground (GND) and red lines are power (VDD). We simultaneously apply an ideal DC voltage source to all nine VDD pads and 0.0 V to GND. The center VDD and GND ports are connected to the power supply pins of the logic gates through the PMOS header switches, as shown in Figure 3.9 and Figure 3.11. Then, using *SPICE* transient analysis, we measure the power-supply fluctuation for various configurations of the 3D IC when the internal logic makes transitions [97].

Figure 3.9 shows the TSV-based 3D IC stacking methodology. Multiple dies are stacked in face-to-back bonding. Figure 3.10 shows the IR-drop results that are related to the voltage fluctuations of the on-chip PDN with a single-layer TSV. The TSV causes an additional voltage drop of 127% and a ground bump of 110% in a single layer.

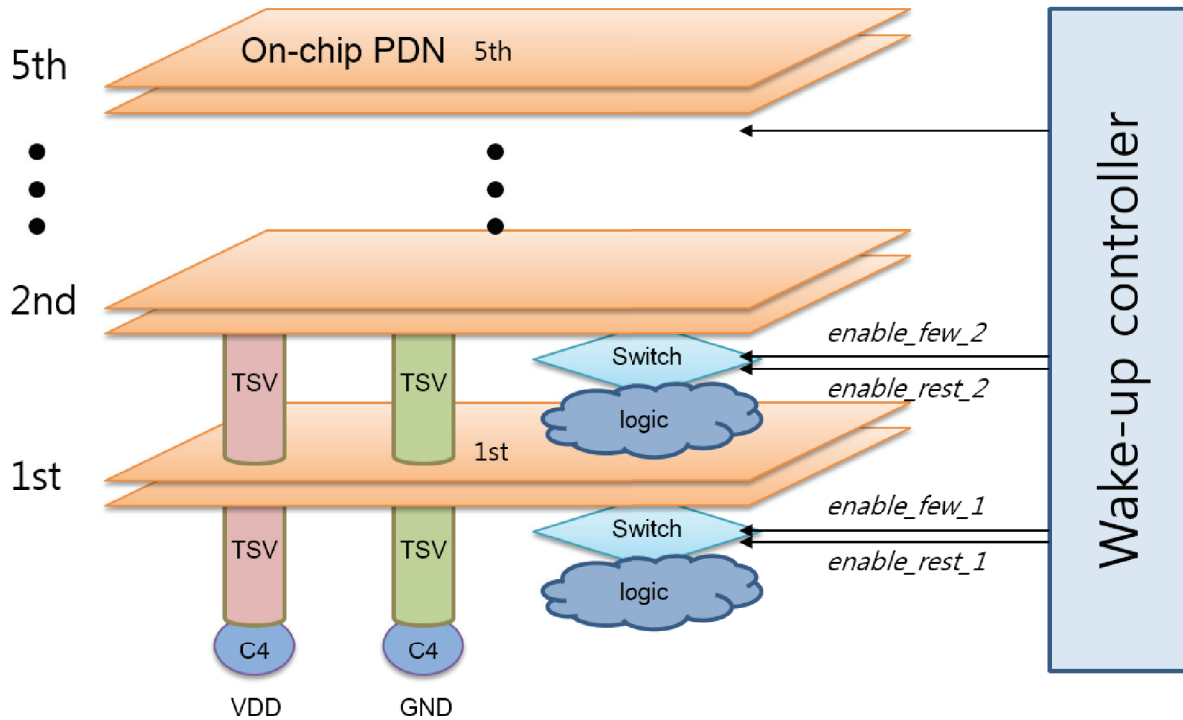


Figure 3.9: 3D IC stacking for a total of five layers with power gating switches and logic in each layer. Face-to-back stacking is used. VDD and GND are supplied from the bottom C4 bumps. The proposed wake-up controller minimizes the wake-up latency of each layer; it controls the interval delay between the *enable_few* and *enable_rest* signals.

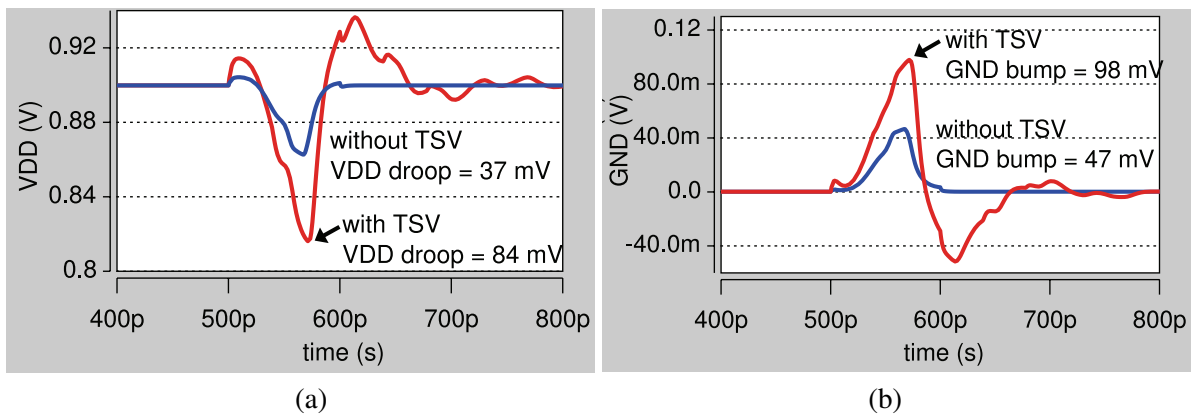


Figure 3.10: VDD drop (a) and GND bump (b) comparison with and without a TSV in a single-layer die. A clock buffer that is 40 times larger than the minimum clock buffer with an input slew of 100 ps is used in this simulation.

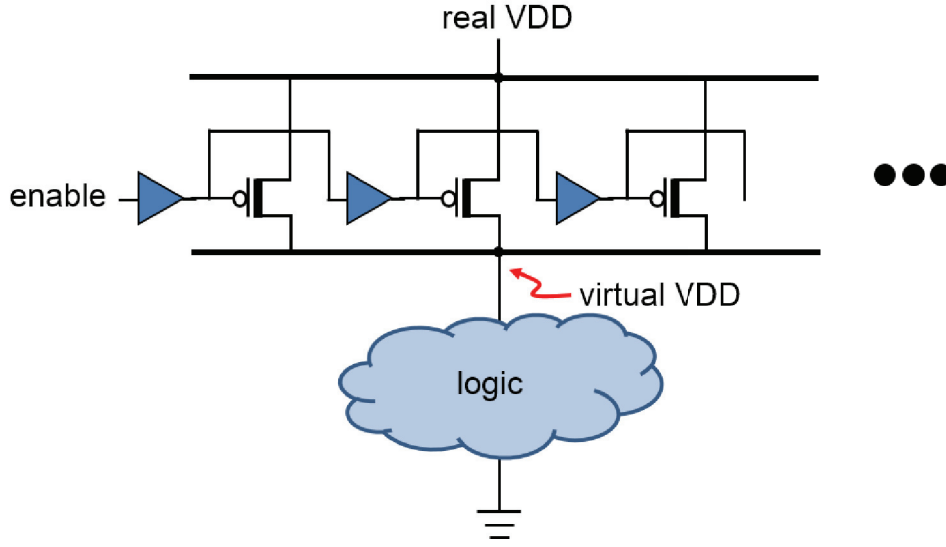


Figure 3.11: On-chip PDN model of one layer of a 3D IC with PMOS header switches for power gating. The enable-low control signal propagates through all of the switches, with additional buffers in the daisy chain.

3.3 Power Gating Design in 3D IC

In this section, we describe a general power gating strategy to apply in 3D IC with our simulation setup. Figure 3.9 shows the face-to-back 3D IC stacking structure with a total of five layers of dies; the PMOS header switches are located in each layer. We assume that the $0.5\text{ mm} \times 0.5\text{ mm}$ die has nine VDD TSVs and nine GND TSVs, as shown in Figure 3.3. Then, current is delivered to the logic gates through the on-chip PDN and TSVs between the layers. For the *SPICE* simulation, we use the predictive technology model (PTM) for 22-*nm* devices [100]. The nominal VDD value is set to 0.9 V. The logic consists of 15,000 clock buffers to mimic the current load in the simulations. Each clock buffer consists of two inverters with $W_n = 4\text{ }\mu\text{m}$ and $W_p = 8\text{ }\mu\text{m}$. Each die consists of three components; on-chip PDN, logic, and additional internal metal wires for PMOS header switches. The internal metal wire capacitance, which is 100 pF is estimated based on the size of the chip and the number of inserted switches. Also non-switching layers are acting as decoupling capacitors through charge sharing. For 3D IC analysis, we stack these die models through TSVs in *S*-parameters and vary the number of stacked layers to identify the impact of multiple layers on the IR drop and in-rush current.

3.3.1 Power gating

The insertion of PMOS headers or NMOS footers to implement a power gating technique is a promising solution for minimizing the standby leakage current. In this study, a PMOS header is used for the power gating switches, as shown in Figure 3.11. The number of power gating switches and logic clock buffers is calculated using the same basis as that used in [96, 79]. When the logic area is 1.635 mm^2 for a 22-nm high-performance (HP) design, a total of 5,516 switches are required. Our representative, scaled die area is 0.25 mm^2 ; thus, the total number of switches we have inserted in our simulation is 843 for each layer. We also set the number of logic clock buffers without scaled die area overhead. HP (or nominal V_{th}) devices are used for the clock buffers, and low-standby-power (LSTP) (or high- V_{th}) devices with a width of $4 \text{ }\mu\text{m}$ are used for the PMOS header switch. The leakage current of one PMOS header switch is $15.8 \text{ pA}/\mu\text{m}$ (or a total of 53.3 nA). The number of switches and their size are selected on the basis of the IR-drop and in-rush current constraints. In our simulation, a maximum IR drop of 5% of nominal VDD (i.e., 45 mV) at the virtual VDD (VVDD) is used. The wake-up time is defined as the interval between the transition of enable signal and the time at which VVDD becomes 95% of the nominal VDD. Thus, when the internal VVDD reaches 0.855 V (95% of nominal VDD), each layer finishes the wake-up sequence.

3.3.2 Single-stage power gating with daisy-chain buffer

During idle (or standby) mode, the PMOS header switches are “off”, and therefore, disconnect VVDD from real VDD to minimize leakage power dissipation through the logic gates. When the logic starts to operate, the “enable” (i.e., enable-low) signal is applied to turn on the PMOS switches. Then, a very large amount of in-rush current simultaneously flows through the switches to charge VVDD to the real VDD value within a short period of time. Therefore, any victim logic that is already operating experiences a significant voltage fluctuation (i.e., IR drop) when the other logic gates (i.e., aggressors) are attempting to turn-on. A larger IR drop in the victim logic is expected if we insert larger switches; however, we need both a certain number of switches and a sufficient current to reduce the wake-up time (i.e., the time for VVDD to reach the real VDD value from zero). The conventional approach to minimizing the IR drop in the presence of a large number of very large switches is to avoid the simultaneous turn-on scenario. There are several techniques to delay the successive switches. The simplest method is a single-stage turn-on scheme that connects daisy-chained buffers for each switch, as shown in Figure 3.11 and Figure 3.12(a). The daisy-chain buffers in each switch continuously delay the enable signal and reduce the IR drop. However, to satisfy the IR-drop constraint, the daisy-chained

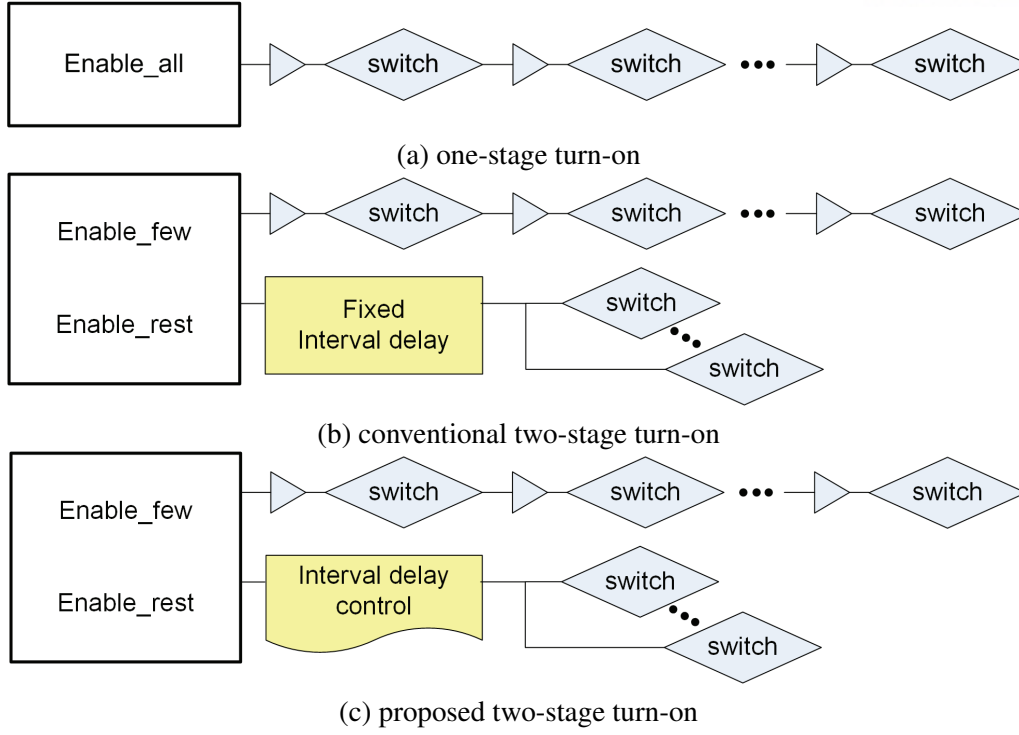


Figure 3.12: (a) Conventional one-stage turn-on scheme with a daisy chain, (b) two-stage scheme using the *enable_few* and *enable_rest* signals, and (c) proposed adaptive two-stage turn-on scheme. The interval between the *enable_few* and *enable_rest* signals is controlled by the wake-up controller (see Figure 3.9) on the basis of victim- and aggressor-layer information.

enable signal significantly increases the wake-up time.

In the following section, we propose a new adaptive two-stage turn-on technique in the 3D IC stack to reduce the wake-up time while satisfying the IR-drop constraint.

3.4 Adaptive Two-Stage Power Gating Strategy in a 3D IC

As shown in the previous section, turning on all switches simultaneously results in a significant voltage fluctuation in the victim layer; a longer turn-on time is required to satisfy the IR-drop limitation in the single-stage daisy-chain method. Thus, multiple-stage enabling techniques in which several bundles of switches are sequentially turned on after a certain interval have been proposed [19, 20, 21, 22, 79]. More stages provide better in-rush current and IR drop control, but have the disadvantage of a longer turn-on time. The two-stage enable technique is believed to exhibit a good tradeoff between the in-rush current and the turn-on time [79]. In this study, we exploit the two-stage enable technique for a 3D IC and propose an adaptive two-stage power gating strategy to minimize the wake-up time based on an understanding of the layer dependencies.

Our adaptive two-stage turn-on architecture is shown in Figure 3.12(c). As shown, first, several switches turn on with the daisy chains to satisfy the IR-drop constraints; all other switches begin to turn on to charge VVDD. The conventional two-stage enable technique results in the worst case of the interval delay (between the *enable_few* and *enable_rest* wake-up signals), which satisfies the IR-drop constraints. However, because there are many wake-up situations in a 3D IC, the (worst-case) same interval delay can waste the wake-up time. For example, a five-core layer has 180 wake-up situations according to the configuration of all layers, and each case requires a different wake-up time to satisfy the constraints.

If we turn on more switches in the *enable_few* signal stage, a larger IR-drop is expected with a shorter wake-up time. In addition, fewer switches in the *enable_few* signal stage require a large interval delay to satisfy the IR-drop constraint. Therefore, selection of the ratio between the few and rest switches becomes important. Analysis of several different few-to-rest ratios has been conducted to determine the optimum switch ratio; the simulation results are shown in Figure 3.13. The worst IR drops and wake-up times are summarized in Table 3.4. The interval between the last *enable_few* signal and the beginning of the *enable_rest* signal is 13.0 ns, assuming that this takes 26 clock cycles in a 2 GHz system. As shown, the two-stage enabling scheme with a ratio of 0.5:9.5 results in a wake-up time reduction of up to 31% (compared to the single-stage method) while satisfying the IR-drop constraint. The 1:9 ratio provides the shortest wake-up time (e.g., 18.4 ns); however, it violates the IR-drop constraint. The size of the daisy-chain buffer inside the switches of the few stage (in the two-stage method) is twice as large as that of the single stage. In the two-stage method, the size of the enable buffers can be increased compared to the single stage method. The reason is that two-stage buffer method loosens the IR-drop constraints; under the relaxed constraints, we can increase the buffer size to minimize the total wake-up time. The optimal buffer size has been obtained from *SPICE* simulations. Although the buffer size gets larger

Table 3.4: Worst-case IR drop and wake-up times for different few-to-rest ratios

power gating method	IR drop (mV)	wake-up time (ns)	wake-up time (normalized)
one-stage	46.4	27.50	1.00
two-stage (0.25:9.75)	51.5	19.50	0.71
two-stage (0.5:9.5)	44.1	19.01	0.69
two-stage (1:9)	46.0	18.40	0.67
two-stage (2:8)	49.0	20.50	0.75

than in the single stage method, the total area becomes smaller in the two-stage method because the daisy-chained enable buffers are only required for switches which are controlled by *enable_few* signal.

The worst-case analysis in the previous section, in which all four dies simultaneously begin to turn on and affect the one victim die, is too pessimistic. In addition, in a 3D IC, if two dies are located far away, the impact on the victim die by the aggressor is smaller than when the two dies are closer. Therefore, if we know the location of the aggressor die and the number of aggressors, an adaptive two-stage turn-on scheme to reduce the wake-up time is possible; we can control the interval between the *enable_few* and *enable_rest* signals according to the configuration of each layer. The worst IR drop versus the location of the aggressor die in a 3D IC is shown in Figure 3.14 and the included table. In this analysis, any layer from the top (5^{th}) to bottom (1^{st}) die is assumed to be a victim layer, and only a single layer (i.e., aggressor) starts to turn on at one time. As shown, a larger IR drop is expected when the aggressors are located in the adjacent layers (i.e., for the i^{th} victim die, the aggressors are in the $i - 1^{th}$ or $i + 1^{th}$ layer). On the other hand, a relatively small IR drop occurs if the aggressor die is far from the victim die. The worst IR drop (38.5 mV) occurs in the 5^{th} layer from an aggressor in the 4^{th} layer with the fixed interval. In addition, the victims in the higher layers have a greater IR drop than the lower layers because higher layers are located far from the VDD source. Thus, placing frequently used or long-running layers (i.e., victim layers) nearby VDD source (or lower stack) can improve a certain amount of wake-up time since lower layers have smaller IR-drop than upper layers.

We can take advantages of this result. There is a tradeoff between the IR drop and the wake-up time, and we can shorten the interval between the *enable_few* and *enable_rest* signals within the 5% VDD IR-drop constraint. There are three conditions in each core layer: (i) running, (ii) sleeping (or power-gated), and (iii) starting to wake up. Already-running layers can be considered victim layers. Power-gated layers can be considered idle layers. Layers that are starting to wake up are considered aggressors. [35] considers only two layer conditions to simplify the layer combinations, which are aggressor (starting to wake up) and victim (running or sleeping). However, running and sleeping (power gated) layers have

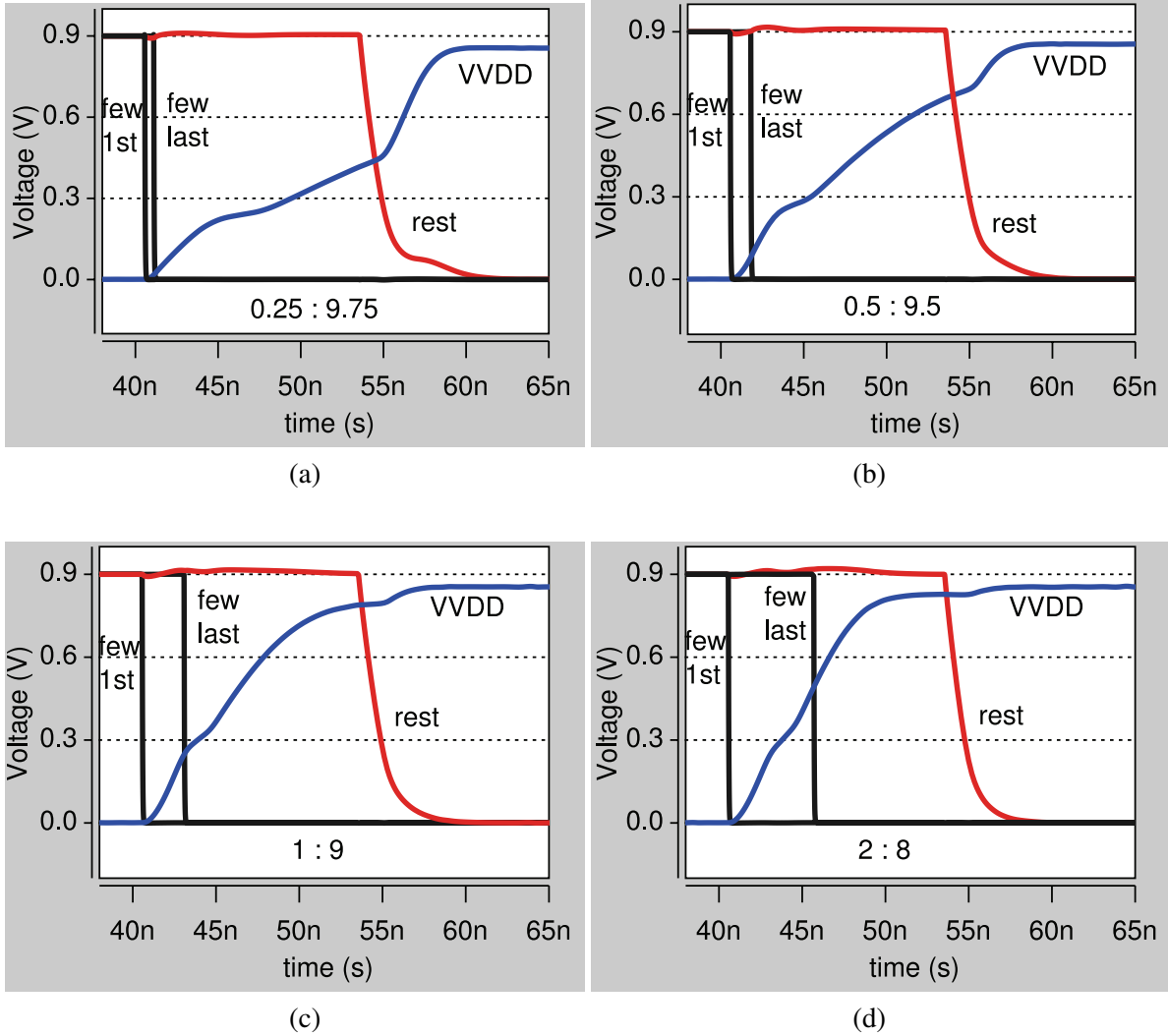


Figure 3.13: The *enable_few* (black line) and *enable_rest* (red line) signals for different few-to-rest ratios. The VVDD waveform is shown by the blue line. The few-to-rest ratios are shown at the bottom of each plot. “Few 1st” means the first turned on *enable_few* switch; “Few last” is the last switch. The ratios are (a) 0.25:9.75, (b) 0.5:9.5, (c) 1:9, and (d) 2:8.

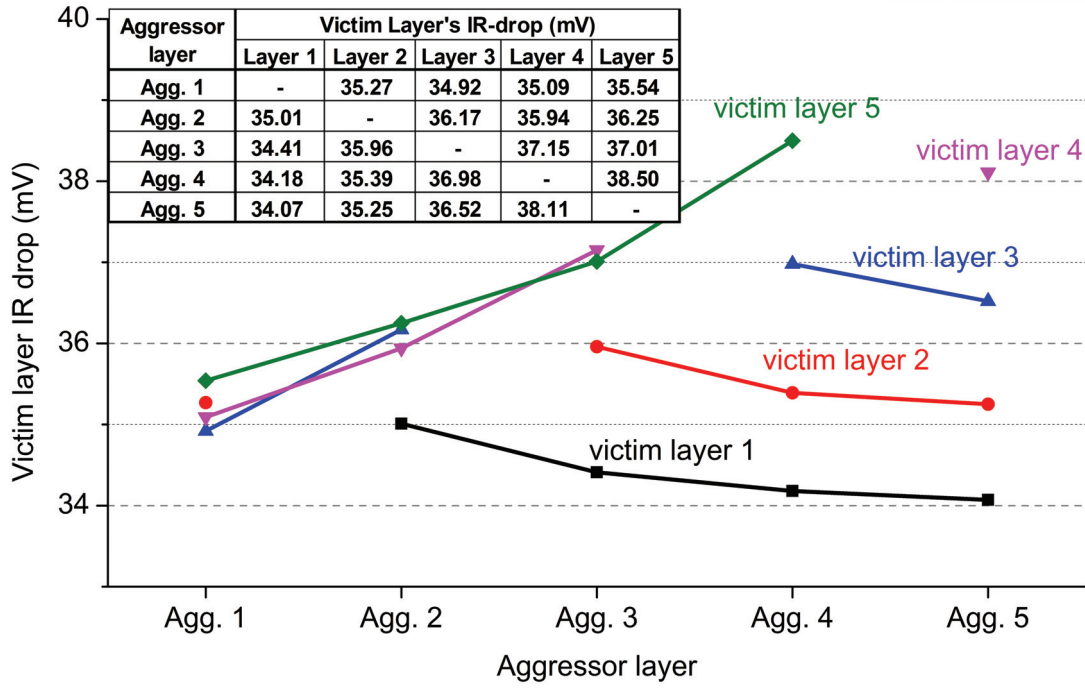


Figure 3.14: IR-drop dependency according to the location of the aggressor and victim layers. The IR drop in the victim layer increases when the aggressors are located in adjacent layers. The absolute values are summarized in the table included in the figure.

different IR drop constraints. The simplification requires a worst-case assumption (e.g., all victim layers are running), and results in a pessimistic wake-up control. In this work, we consider the three conditions for each layer to make the analysis precisely.

In the following section, we propose a simple algorithm to find the optimum intervals between *enable_few* and *enable_rest* signals for all possible layer configurations.

3.5 Optimization of Adaptive Interval

In our proposed adaptive method, the maximum allowable interval for each aggressor layer, considering all possible operating conditions (i.e., idle, aggressor, or victim), can be obtained from a simple algorithm, as illustrated in Algorithm 1. As shown, all layers are initialized using the idle state (Line 1), and an aggressor combination is selected from among all layers; then, possible victim sets are selected for each aggressor combination (Lines 7-11). Then, the proposed IR-drop analysis for the adaptive power gating environment is conducted while sweeping the interval between the *enable_few* and *enable_rest* signals and decreasing from the maximum interval (which is obtained from the conventional two-stage power gating method) (Lines 12-16). Finally, if the worst IR drop value in the victim layers is smaller than the constraint, the interval will be decreased until the interval time is zero (Lines 17-18). However, when the worst-case IR-drop is worse than the constraint, it will return to the previous interval (i.e., the not violating value) and provide wake-up time information for each aggressor (Lines 19-22). The total number of aggressor and victim combinations for a 3D IC of N layers can be expressed as

$$\sum_{i=1}^{N-1} {}^N C_i \cdot \left(\sum_{j=1}^{N-i} {}^{N-i} C_j \right),$$

where i is the number of aggressors and j is the number of victims. For example, when the number of layers is five, there are 180 possible combinations in total

$$({}^5 C_1 \cdot ({}^4 C_1 + {}^4 C_2 + {}^4 C_3 + {}^4 C_4) + {}^5 C_2 \cdot ({}^3 C_1 + {}^3 C_2 + {}^3 C_3) + {}^5 C_3 \cdot ({}^2 C_1 + {}^2 C_2) + {}^5 C_4 \cdot ({}^1 C_1)).$$

Figure 3.15 shows three examples of the interval-finding algorithm result. For example, in (a), the bottom layer (1st layer) is an aggressor, the top layer (5th layer) is a victim, and the others (2nd, 3rd, and 4th layers) are in the idle state. In this case, there is only one aggressor, and the aggressor layer is far from the victim layer; thus, a 0 ns interval is sufficient to satisfy the IR-drop constraint. (b) is one of the examples that has an interval between 0 ns (minimum) and 12.5 ns (maximum). In this case, aggressors are in both the 2nd and 4th layers, the victims are in the 1st and 5th layers, and an idle layer is in the 3rd layer. This case requires a 6.5 ns interval for the shortest wake-up time while also satisfying the 5% IR-drop constraint (above 0.855 V). The case in (c) is the voltage profile of the worst-case combination. The 4th layer is the only victim layer and all other layers are operating as aggressors. In this case, the maximum interval (e.g., 12.5 ns) is required so as to satisfy the IR-drop constraint.

By considering the idle state of the layer combination, the proposed adaptive method effectively

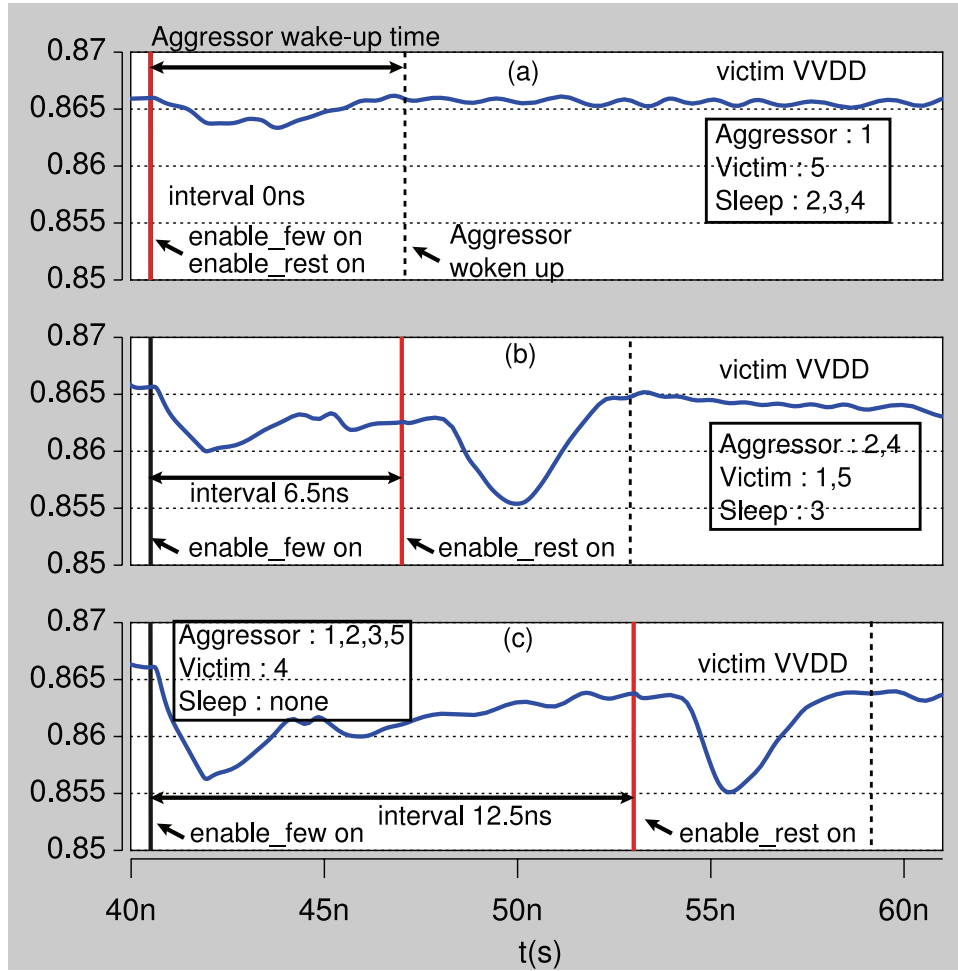


Figure 3.15: Voltage drop profiles of the three combinations. The blue line is the virtual VDD of the worst victim layer. The vertical black and red lines represent the starting time of the *enable_few* signal and the *enable_rest* signal, respectively. The vertical dotted line is the worst wake-up time among aggressors. (a) The best IR-drop combination, which has a 0 ns interval between the *enable_few* and *enable_rest* signals. (b) The case of a 6.5 ns interval. (c) The worst-case IR-drop combination, which requires a 12.5 ns interval.

ALGORITHM 1: Interval finding algorithm for the proposed adaptive method

input : N_{layers} \leftarrow total number of layers, the worst-case IR-drop constraint
output: Adaptive interval time of each case

- 1 Initialization: all layers are in idle state;
- 2 $interval_{MAX}$: maximum interval \leftarrow conventional 2-stage constant interval method;
- 3 N_{agg} : number of aggressors;
- 4 N_{vic} : number of victims;
- 5 $N_{non-agg}$: number of non-aggressors (victim or idle);
- 6 ${}^N C_K$: N choose K combination;
- 7 **for** $N_{agg} \leftarrow 1$ **to** $N_{layers} - 1$ **do**
- 8 Aggressor set = ${}^{N_{layers}} C_{N_{agg}}$;
- 9 Non-aggressor layers = layers \notin Aggressor set;
- 10 **for** $N_{vic} \leftarrow 1$ **to** $N_{layers} - N_{agg}$ **do**
- 11 Victim set = ${}^{N_{non-agg}} C_{N_{vic}}$;
- 12 **for** $interval \leftarrow interval_{MAX}$ **to** 0 **do**
- 13 Update layer status;
- 14 IR-drop analysis with new interval;
- 15 Find the worst-case IR drop among all victim layers;
- 16 Find the wake-up time of all aggressor layers;
- 17 **if** *worst-case IR-drop in victims* \leq *IR-drop constraint* **then**
- 18 continue with a smaller interval;
- 19 **else**
- 20 interval \leftarrow previous interval;
- 21 wake-up time \leftarrow previous wake-up time;
- 22 exit;
- 23 **end**
- 24 **end**
- 25 **end**
- 26 **end**

reduces the average wake-up time, relative to the results in Section IV. As shown in Figure 3.16, many cases require zero interval delay because layers in the idle state do not affect the IR drop of the victim layers. Therefore, an approximately 42.7% reduction in the wake-up time compared to the conventional two-stage method can be achieved. By investigating these optimization results, we can build a look-up table (LUT) that has the minimum safe interval time between the *enable_few* and the *enable_rest* signals for every layer configuration. Then, a wake-up controller, as shown in Figure 3.9, can use the information of the LUT to turn on all aggressors without violating the voltage drop constraint of any victim layer. The wake-up controller has layer configuration and decide layer status. Thus, when some layers wake-up, the controller refers to the LUT and sends the proper interval delay for achieving the shortest wake-up time without violating IR-drop constraint. The LUT provides information about the intervals (between few and rest signals) for all scenarios, real-time monitoring is not required in the proposed architecture.

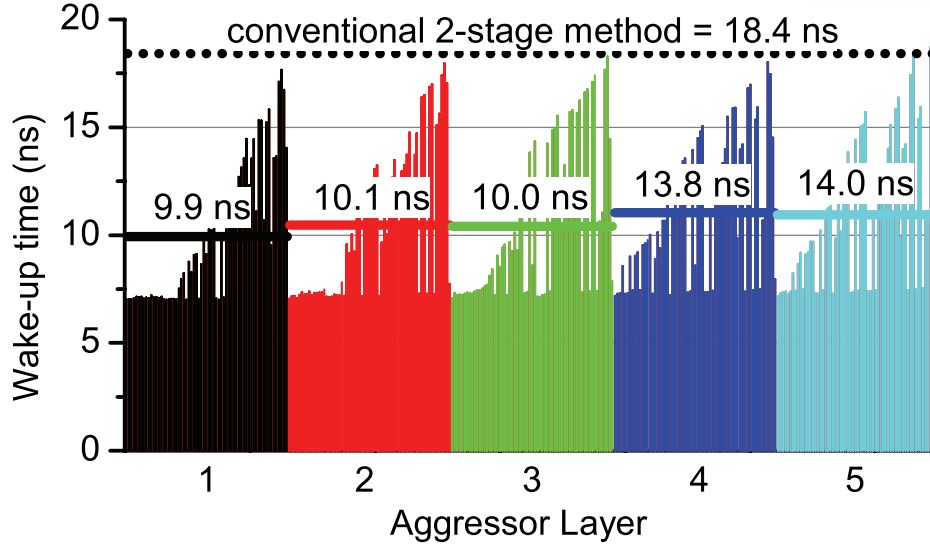


Figure 3.16: Adaptive wake-up time profile for each aggressor layer based on the layer configurations, IR-drop requirement of the victim layer, and idle layers in the 3D IC. A total of 180 cases of layer combinations for the wake-up situation. Overall average wake-up time (of each aggressor layer) is reduced because it considers the idle state, which is ignored in the simulations in Section 3.4.

3.6 Effect of Tapered TSV Structure

For TSV simulations in the previous sections, we have assumed uniform TSVs of $10\text{-}\mu\text{m}$ in diameter for power delivery between all layers. However, there is a vertical layer dependency in a 3D IC because of the difference in the distance of layers from a power source (i.e., C4 bump). As shown in Figure 3.17, the VDD/GND sources and C4 bumps are placed at the bottom. Therefore, TSVs that are located near the source (e.g., TSVs between 1^{st} and 2^{nd}) have a larger current demand than TSVs in upper layers (e.g., TSVs between the 4^{th} and 5^{th} layers) [36].

It is apparent that the larger TSV is better for the satisfying IR-drop constraint because of its smaller intrinsic resistance. However, the enlargement of TSV size reduces the area for logic placement and routing because a keep-out zone (KOZ), surrounding the TSV, of a certain size is required; all logic cells must be placed outside the KOZ to avoid the influence of TSV-induced stress [37]. In addition, the KOZ is proportional to the square of the TSV's diameter.

To investigate the effect of the tapered TSVs on the power gating, we vary the scale of the tapering for the TSV in three cases, as shown in Table 3.5; in case 1, we maintain the total TSV sizes (and KOZ requirements) by reducing the diameters of the upper-layer TSVs and increasing those of the lower-layer TSVs (relative to the nominal diameter), such as $14\text{ }\mu\text{m}$, $12\text{ }\mu\text{m}$, $10\text{ }\mu\text{m}$, $8\text{ }\mu\text{m}$, and $6\text{ }\mu\text{m}$ diameters from the bottom (1^{st}) to the top (5^{th}) layers. In case 2, the diameter of the TSV is gradually enlarged

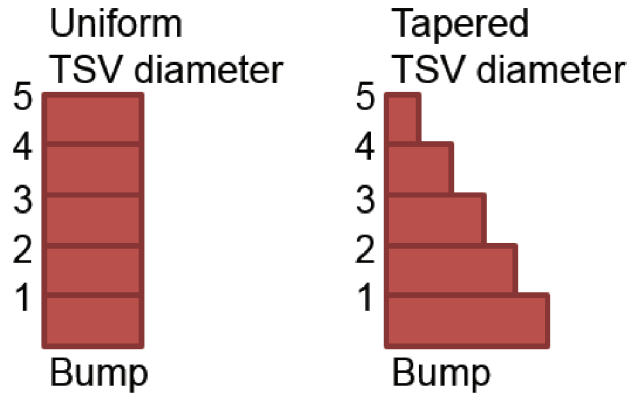


Figure 3.17: Uniform TSV (left) and tapered TSV (right) examples. Larger TSVs satisfy larger current demands. Therefore, we place larger TSVs near the VDD source (i.e., lower layers) and smaller TSVs in the upper layers.

Table 3.5: Conventional and proposed wake-up times with different tapering cases

TSV tapering case diameter (1 st to 5 th) (μm)	wake-up time (ns)			wake-up reduction from tapering	overall wake-up reduction
	conventional w/o optimization	conventional w/ optimization	proposed w/ optimization		
uniform (10, 10, 10, 10, 10)	18.41	18.41	10.56	42.6%	42.7%
case 1 (14, 12, 10, 8, 6)	18.09	15.37	8.71	43.3%	51.9%
case 2 (18, 16, 14, 12, 10)	17.76	10.61	7.34	30.8%	58.7%
case 3 (50, 40, 30, 20, 10)	17.59	9.61	7.11	26.0%	60.0%

(from the nominal diameter) by 2 μm from top to bottom layer. In case 3, we taper the TSVs using 10- μm steps from the nominal diameter.

The extracted RLGC values of the various TSV diameters are shown in Figure 3.18. As shown in the figures, the IR drop becomes smaller as the diameter of the TSV becomes larger because the series resistance (R) becomes smaller. Although the larger conductance (G) may increase the high-frequency insertion loss, the IR drop is dominated by the series resistance at low frequencies (up to 500 MHz).

The wake-up time profiles of the three tapered TSV cases are represented in Figure 3.19. By comparison with the uniform TSV, shown in Figure 3.16, all of the tapered TSV cases result in shorter and more balanced wake-up times among all possible layer configurations. This result indicates that larger TSVs successfully reduce the IR drop in critical places such as the bottom layer. The wake-up time in the conventional method and the proposed adaptive method for various TSV structures are compared and summarized in Table 3.5. The first column shows results with a same interval obtain from the uniform case and the results in the second column are with modified intervals obtain from each tapering case

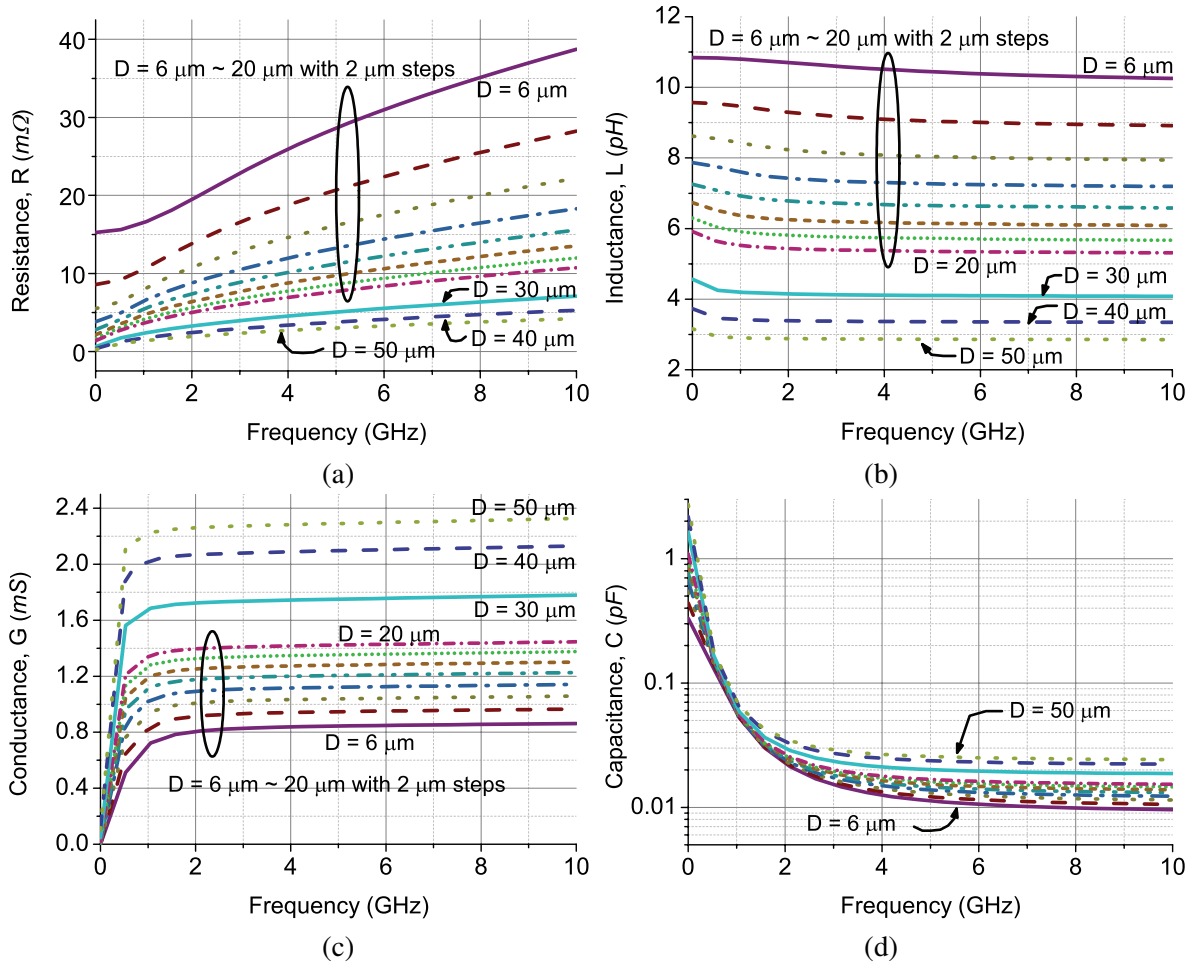


Figure 3.18: Frequency-dependent extracted RLGC [(a) resistance, (b) inductance, (c) conductance, and (d) capacitance] values of TSVs in different diameters.

without proposed optimization process. The third column is the results when the proposed optimization is used in each tapering case. The reduction of the tapering optimization is shown in the fourth column, and the overall benefit by the proposed method is shown in the last column. As expected, the tapered TSVs provide further wake-up time reduction from uniform TSV both in the conventional and the proposed power gating methods. As shown in Table 3.5 and Figure 3.19, tapering TSV itself provides some wake-up time reduction for the conventional method. Further wake-up time reduction (up to 43% or from 15.37 *ns* to 8.71 *ns* for the first aggressor layer in tapering TSV case 1) is possible by applying our proposed algorithm. More reduction in wake-up time is possible by using larger size of tapering (e.g., case 2 and 3) for conventional 2-stage method. However, the proposed optimization still provides reduction at least 26% in average. In addition, the benefit due to the proposed method is larger than that by the conventional method (e.g., 5% vs. 33% more reduction for case 3). Tapering case 3 shows the fastest average wake-up time performance due to the large overall TSV sizes. However, in this case, a larger area penalty is expected than in other cases. The normalized wake-up time and the KOZ are plotted in Figure 3.20. Case 1 shows up to 18% reduction in the wake-up time compared to the uniform TSV case with the same total TSV diameter and minimum KOZ penalty (i.e., 8%). As a result, the tapering approach of case 1 is a good solution for saving wake-up time with minimal penalty to the logic area. Tapering case 2 provides a 12% shorter wake-up time than case 1 with almost twice the KOZ penalty. Therefore, case 2 can be chosen as a tradeoff between area and wake-up time. For case 3, however, the benefit in terms of wake-up time is relatively small and there is a significant penalty in the KOZ. Therefore, a TSV size that is too large (in diameter) than required is not effective in the tradeoff between performance and the area penalty.

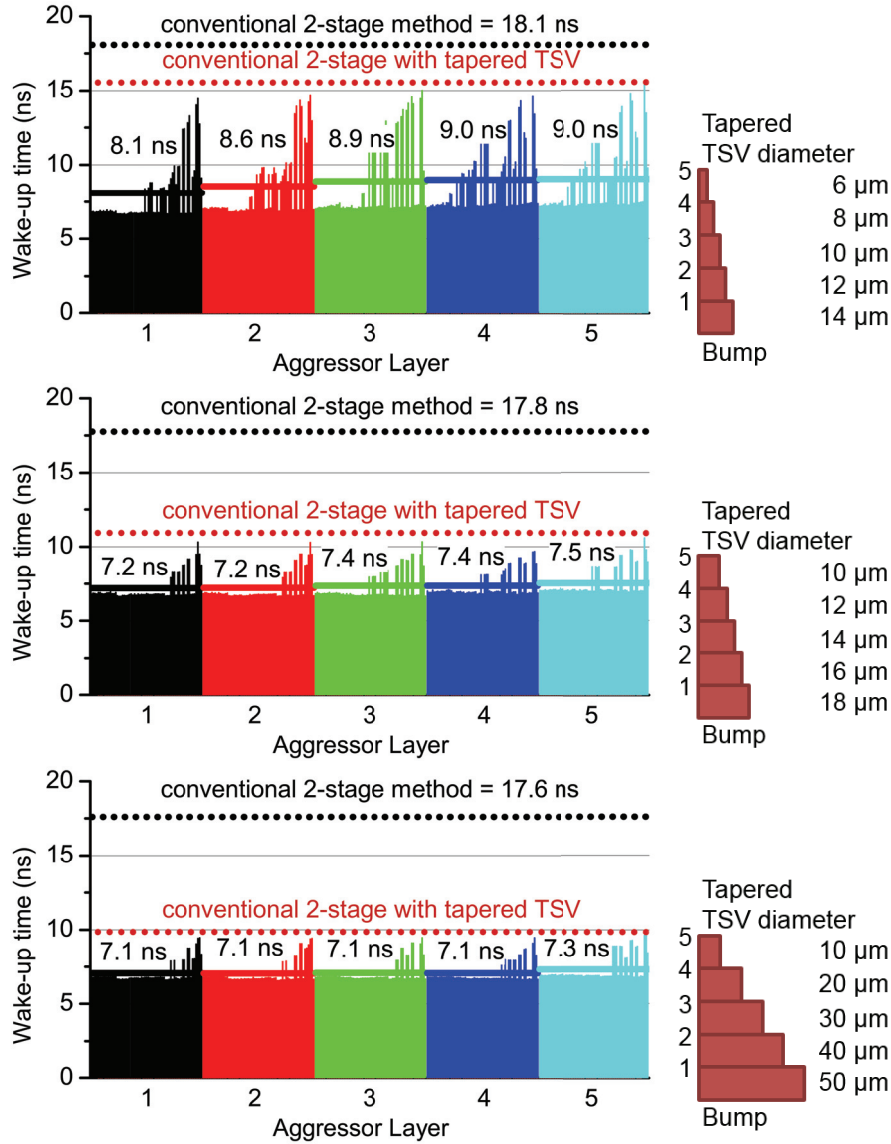


Figure 3.19: Wake-up time profile with tapered TSV cases 1, 2, and 3. The black dot line is average wake-up time of conventional 2-stage method and the red dot line is average wake-up time of conventional 2-stage method with tapered TSV only. The results show not only an improved wake-up time but also more balanced values among layer configurations (i.e., victim, aggressor, and idle) by the tapered TSV than the uniform TSV.

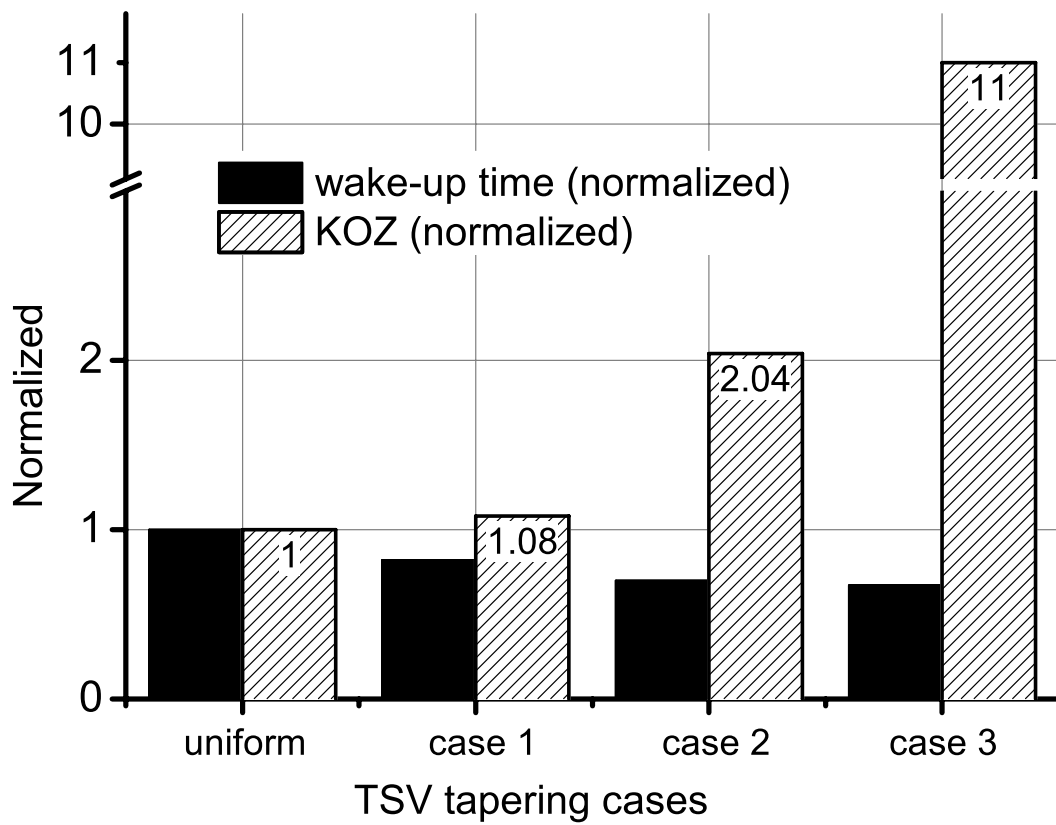


Figure 3.20: Normalized bar graph of the wake-up time when using the proposed method and KOZ with various tapering cases.

3.7 Conclusions and Future Directions

In this chapter, we first investigate the voltage noise of multilayer 3D IC with a proposed analysis methodology of PEEC-based PDN and frequency-dependent TSV models. We find that TSV significantly affects voltage fluctuation in the PDN. Moreover, the inductance component and frequency-dependent TSV models are vital for accurate understanding of the power integrity of the 3D IC. Second, we propose and investigate the multi-paired and wire-added on-chip PDN structure to reduce the voltage noise of PDN in 3D IC. The multi-paired structure effectively reduces IR-drop by approximately a maximum 21%, and as the number of layers increases, the advantage decreases. Additionally, we analyze the correlation between wire space and voltage noise. Simulations based on our proposed method reveal that shorter wire space, by adding VDD and GND wire pairs, reduces $L di/dt$ noise.

We also investigated a power gating technique for a 3D IC with a PEEC-based on-chip PDN and frequency-dependent S -parameters of TSVs. Because there are TSVs between layers, a greater voltage drop is expected in a 3D IC; we have analyzed the in-rush current and the IR drop when the power gating switches turn on. Our analysis shows that a larger voltage drop occurs in the layer farthest from the power supply because of the accumulated drop through the TSVs and on-chip PDN in each die. A two-stage wake-up scheme have been investigated in the PDN of the 3D IC; a few-to-rest switch ratio of 0.5:9.5 provided a wake-up time reduction of 31% compared to the simple single-stage power gating method while satisfying the IR-drop limit. The multiple-die 3D IC analysis indicates that a greater IR drop is expected when the victim and aggressor dies are close, and that a smaller IR drop occurs when they are far away. Therefore, an additional wake-up time reduction can be achieved by controlling the interval between the enable signals (i.e., *enable_few* and *enable_rest*) according to the configuration (e.g., location and operation status) of each layer in the 3D IC structure. An average wake-up time reduction of 28% can be achieved by employing the proposed adaptive method, as compared to the conventional two-stage scheme. This corresponds to a reduction of approximately 50%, as compared to the single-stage method.

Moreover, we have suggested an interval finding algorithm for the proposed adaptive method that considers three possible operating states for each layer. Up to 43% wake-up time reduction compared to the conventional two-stage method can be achieved by employing the optimized interval algorithm. In addition, tapered TSVs have been analyzed in the proposed power gating methodology. Simple tapering (e.g., case 1) provides additional 18% wake-up time reduction from the uniform TSVs with minimal area overhead. In this research, we use a fixed regular PDN structure to obtain the optimal interval solution by applying our proposed algorithm. The entire flows are general, thus any on-chip PDNs and TSV structures including the tapered TSV can be exploited. Hence, the main contribution

of ours is to find an optimal wake-up interval and tapered TSVs by PI analysis in multi-layered 3D IC design.

Our ongoing work seeks to analyze the IR drop and implement a power gating strategy for heterogeneous 3D IC layers. In addition, we will analyze other powering mode scenarios that perform on different places on the chip. More broadly, our research will model the minimum wake-up latency based on the layer status at each location and apply the modeled latency to 3D IC power gating.

3.8 Acknowledgments

Chapter III is in part a reprint of “Analysis and Reduction of Voltage Noise of Multi-layer 3D IC with PEEC-based PDN and Frequency-dependent TSV Models”, *Proc. IEEE International SoC Design Conference*, 2014; “Novel Adaptive Power Gating Strategy of TSV-based Multi-layer 3D IC”, *Proc. IEEE International Symposium on Quality Electronic Design*, 2015; “Analysis and Reduction of the Voltage Noise of Multi-layer 3D IC with Multi-paired Power Delivery Network”, *IEICE Electronics Express* 14(18) (2017); and “Novel Adaptive Power Gating Strategy and Tapered TSV Structure in Multi-layer 3D IC”, *ACM Transactions on Design Automation of Electronic Systems* 21(3) (2016).

I would like to thank my coauthors Professor Youngmin Kim, Professor Seokhyeong Kang and Professor Ki Jin Han.

Chapter IV

Power Delivery Pathfinding for Emerging Die-to-Wafer Integration Technology

The semiconductor industry has enjoyed tremendous growth and innovations over the past 50 years, in large part due to the self-fulfilling prophecy of Moore's Law. With foundry 7nm products in high-volume production this year, only a few feasible technology nodes remain to potentially deliver PPAC (power, performance, area, cost) benefits from transistor scaling. To address this upcoming challenge, the past decade has seen 3D IC stacking technologies emerge as the main hope for future scaling of integration, area footprint and design performance / power envelope. However, conventional *packaging-driven* 3D IC integration technologies with TSVs are limited by TSV size and pitch, which constrains achievable vertical integration density [38].

In this chapter, we present an efficient pathfinding methodology for PDN design of emerging D2W-based designs. Our proposed solution comprises a set of models that provide early feasibility and QoR tradeoff studies of various PDN designs. More specifically, we build an IR drop model to predict the worst IR (WIR) drop of a given PDN configuration. To comprehend the effect of a given PDN solution on overall design QoR, we also develop a routability model which predicts the routability of a design given a PDN configuration. Putting these elements together, our pathfinding methodology first filters out PDN configurations based on a given design's prescribed IR drop limits; then, the routability model is used to identify the IR drop feasible PDN configuration(s) that offer best routability. We thus obtain a high-quality, satisfying PDN solution that is "optimal" in the sense of both predicted IR drop and estimated routability within our modeled PDN design space. The PDN solution offers direct benefits to design QoR and ease of implementation. Our main contributions are summarized as follows:

- We study the impact of VI density on design routability and build a VI-aware routability model.
- We propose an interface to properly combine IR drop analysis of PDN configurations and corresponding impact on routability.
- On a 28nm design, our model identifies a PDN in the top-3 out of 256 possibilities.
- To the best of our knowledge, we are the first to propose such a pathfinding methodology to identify optimal PDN configurations for D2W-based designs.

4.1 Related Work

Several design methodologies using existing commercial 2D CAD tools have been proposed for physical implementation of gate-level 3D ICs [40][47][49][51][48]. The Shrunk2D (S2D) flow [49][51] performs gate-level 3D IC implementation, while the subsequent Cascade2D flow implements both gate-level and block-level monolithic 3D IC [40]. (Cascade2D focuses on monolithic 3D (face-to-back, or F2B) and does not support F2F bonding technology.) Recently, a commercial-quality F2F-bonded 3D IC implementation flow Compact-2D (C2D) has been proposed [48]. These 3D implementation flows leave open the issue of power delivery and its management along with routability in a given BEOL context.

A number of power delivery analyses in gate-level 3D ICs have also been published. Panth et al. [50] propose a PDN-centered tier-partitioning technique that considers the IR drop vs. thermal tradeoff in monolithic 3D IC. Samal et al. [54] have analyze full-chip impact of PDN designs in monolithic 3D ICs using real design benchmarks. Optimized 3D PDN design configurations (limited to six categories, depending on design characteristics) are compared across power, performance, IR drop and wirelength metrics in different technology nodes. However, design-specific PDN choices at the “Pareto frontier” of IR drop vs. routability are not addressed, as this would require exploration of PDN structures with degrees of freedom on each metal layer. Chang et al. [41] develop a system-level PDN model, along with static as well as dynamic frequency and time domain analyses. 2D and 3D ICs with extracted equivalent RLC parasitics are compared using a single baseline PDN structure. However, the focus is on dynamic rail analysis with frequency-related environmental differences (e.g., decap insertion) rather than optimization of the PDN specification.

For congestion estimation of BEOL, [46] describes the calculation of intrinsic routing capacity for a given routing resource; the method normalizes routing capacity at each edge of global routing cells (*gcells*) to original capacity, and considers minimum width and spacing on a per-layer basis. Thus,

calculated routing capacity is sensitive to how obstacles such as PDN stripes, block each layer. The “PROBE” approach of [45] gives a methodology to rank BEOL stack options according to an intrinsic routing capacity; our work below uses a routability characterization technique from [45]. Additional works have studied the issue of vertical cuts (interconnect demands) in gate-level 3D IC implementation - e.g., attempting to maximize the benefits of 3D ICs by increasing the number of monolithic inter-tier vias (MIVs) or face-to-face VIs [48][49]. Peng et al. [53] note that as the number of vertical cuts increases, inter-die coupling capacitance increases in 3D IC; this can significantly affect power and signal integrity in F2F bonded ICs.

The need for PDN pathfinding in 3D IC arises because power/ground delivery is far from “free”: in the D2W regime, there are TSV and routability impacts, as well as a need for the PDN solution to support delivery of power/ground and signal through inter-tier VIs. The number of VIs is a significant determinant of power and signal integrity, in light of routing congestion and IR drop. This is in contrast to PDNs in 2D ICs, which are generally less sensitive to signal routing congestion on upper metal layers. If the total number of VIs is high relative to the total number of nets (i.e., #VIs-to-#nets ratio), this means that the number of 3D nets across the VIs located on the top metal layer is also relatively high. Therefore, it is also essential to consider the impact of VIs (induced by a given design partition across tiers) when designing a PDN. While previous studies of 3D IC implementation have illuminated many aspects of partitioning, place-and-route and power delivery, typically only a very limited PDN solution space is considered. Our proposed pathfinding methodology attempts to fill this gap with its explicit consideration of both IR-drop constraint and place-and-route feasibility.

4.2 Methodology

A dense PDN can be obstacles to signal routing, and affects the feasibility of the VI placement in the top layer in 3D ICs or causes an increase in wirelength due to many detours. To obtain an optimal PDN solution, we propose a power delivery pathfinding flow, which explores possible PDNs for a given design with a WIR drop requirement. To effectively and efficiently explore the PDN design solution space, we develop a WIR drop model and a routability model to filter and rank possible PDN designs.

4.2.1 Power delivery pathfinding flow

We define the *power delivery pathfinding* problem as follows.

Power Delivery Pathfinding Problem. Given a placed circuit design where cell locations and VI locations are known, provide a PDN design that (a) meets the IR drop limit requirement and (b) has the best routability.

Inputs: A placed design, VI locations and BEOL stack.

Output: A PDN with best routability that meets the IR drop requirement.

Constraints: WIR drop and technology constraints (e.g., minimum width, minimum spacing, etc.).

Figure 4.1 illustrates our power delivery pathfinding flow. Given a placed design, our goal is to find the optimal PDN within the PDN design solution space, which is enumerated by technology constraints (width, space and pitch) of each power metal stripe. For the enumerated PDN designs, we apply the IR drop model to predict their WIRs, and find PDN designs which satisfy the WIR requirement. We then use our routability model to rank PDN designs based on their routability. Besides the PDN variables including metal width, spacing and pitch, we also consider utilization and VI density of the design in the routability model, so as to comprehend the competition in routing resources between PDN and signal routing. Based on the IR drop and routability models, our flow returns an optimal PDN design which satisfies the WIR constraint, and has the best routability. This PDN solution will provide the highest probability of a clean 3D IC implementation.

4.2.2 PDN design knobs

To explore PDN design space, we use PDN design knobs described in Table 4.1. Circuit design-independent knobs include width, space and pitch size of metal stripe as shown in Figure 4.2. Combinations of these knobs must satisfy the design rule constraints of the given technology library. For a given 3D IC design, we use the number of cell instances, row utilization and VI density as circuit design-dependent knobs.

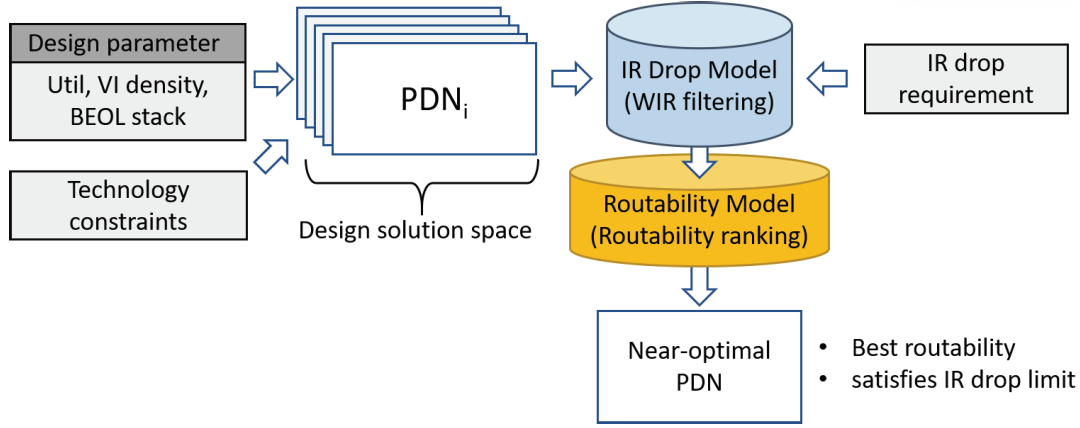


Figure 4.1: Model-based PDN pathfinding flow which gives the optimal PDN design considering both IR drop requirement and routability requirement.

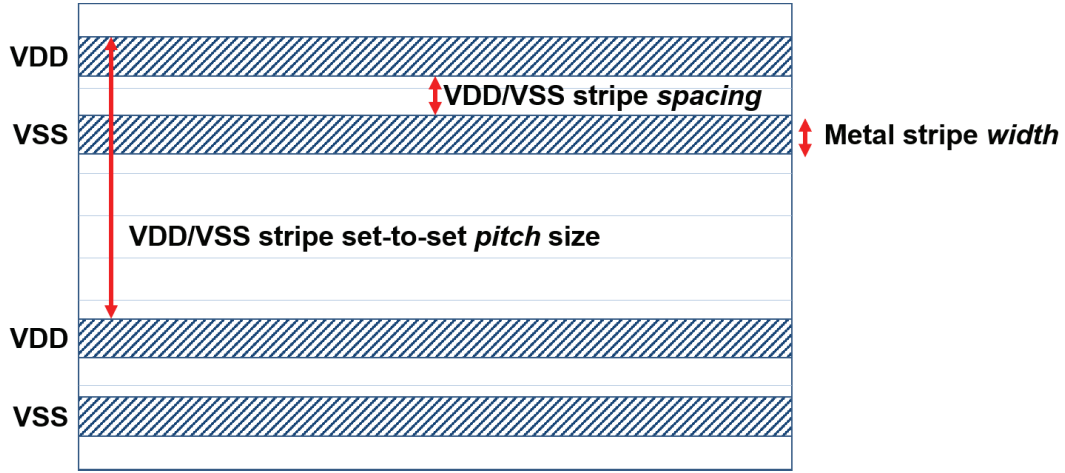


Figure 4.2: Illustration of circuit design-independent PDN design knobs.

4.2.3 WIR & routability modeling

We use nonlinear learning-based algorithms such as multivariable linear regression, and multivariate adaptive regression splines (MARS) [44] to build regression models for both WIR and routability.

WIR Modeling. We build a model to predict the WIR drop for a given PDN design. We use the circuit design-independent knobs as inputs of the WIR model. As mentioned in Section 4.2.1, WIR model is built to prune the PDN design solution space by the WIR requirement. Figure 4.3(a) illustrates the WIR modeling flow. We perform static IR analysis with RedHawk [64] to collect WIR data for various PDN designs. We then model WIR with the aforementioned modeling techniques. Finally we use hybrid surrogate modeling to build a combined model for WIR assessment.

Routability Modeling. We build a model to predict the routability for a given circuit design with

Table 4.1: PDN design knobs.

Circuit design-independent knobs	
Metal stripe <i>width</i> (w)	Width of PDN stripe for each layer
VDD/VSS stripe set-to-set <i>pitch</i> size (p)	Set-to-set distance of VDD/VSS PDN stripe for each layer
VDD/VSS stripe <i>spacing</i> (s)	Spacing between VDD/VSS PDN stripe for each layer
Circuit design-dependent knobs	
#Instances	Total number of instances of circuit in one tier in 3D IC
Utilization	Row utilization of circuit
VI density	The ratio of the number of VIs to the number of nets

pre-routed PDN. Similar in spirit to PROBE [45], we measure the routability of a PDN design by the maximum cell swap count, K *threshold* (K_{th})¹ before exceeding a pre-defined design rule violation (DRV) threshold. In order to train the routability model of PDN, uniform cell placement is needed for gradually increasing routing difficulty as K value increases. Figure 4.4(a) illustrates the mesh-like placement in PROBE. We use 3-input AOI cell for mesh-like placement, and the inputs and output of each cell are connected.

A higher K_{th} value implies that the given PDN design has better routability, i.e., more routing capacity. To understand the impact of VIs on routability of a given PDN in a 3D IC case, we extend the mesh-like placement with connections from cell pin to VI pin on the top metal layer as shown in Figure 4.4(b). We fix the location of VI pins during random neighboring cell swapping. The number of VIs is determined by VI density as the input parameter, and VI density is defined as $\#VIs/\#nets$. The VIs are placed on the top metal and VIs do not overlap the PDN. Note that to implement routing by a commercial 2D P&R tool during the experiment, the VIs in the routability model are placed as I/O pins.

Each VI is connected to the net of the nearest output pin of cells. Figure 4.3(b) illustrates the routability modeling flow. We perform PROBE-like routability analysis [45] with Innovus [63] to collect K_{th} data for various PDN designs.

We also apply model-based calculation methods in addition to design knobs to predict routability of PDNs. A routing capacity score is used for each BEOL layer in the same way as in [46]. We calculate the routing capacity score for each *gcell* edge based on a *gcell* grid from Innovus, and then the total routing capacity score for each layer is obtained by adding all the scores of edges.² We then model K_{th}

¹The K value indicates number of neighbor-swaps normalized to the total number of instances for a given placement, and K_{th} is the minimum K value when routing fails. Following [45], we define routing failure as $\#DRVs > 150$.

²Readers are referred to reference [46] for the detailed calculation method.

value with the aforementioned modeling techniques. By combining several models (multivariable linear regression and MARS), we achieve a hybrid surrogate model to assess the routability of PDN design. Model validations are discussed in Section 4.3.4.

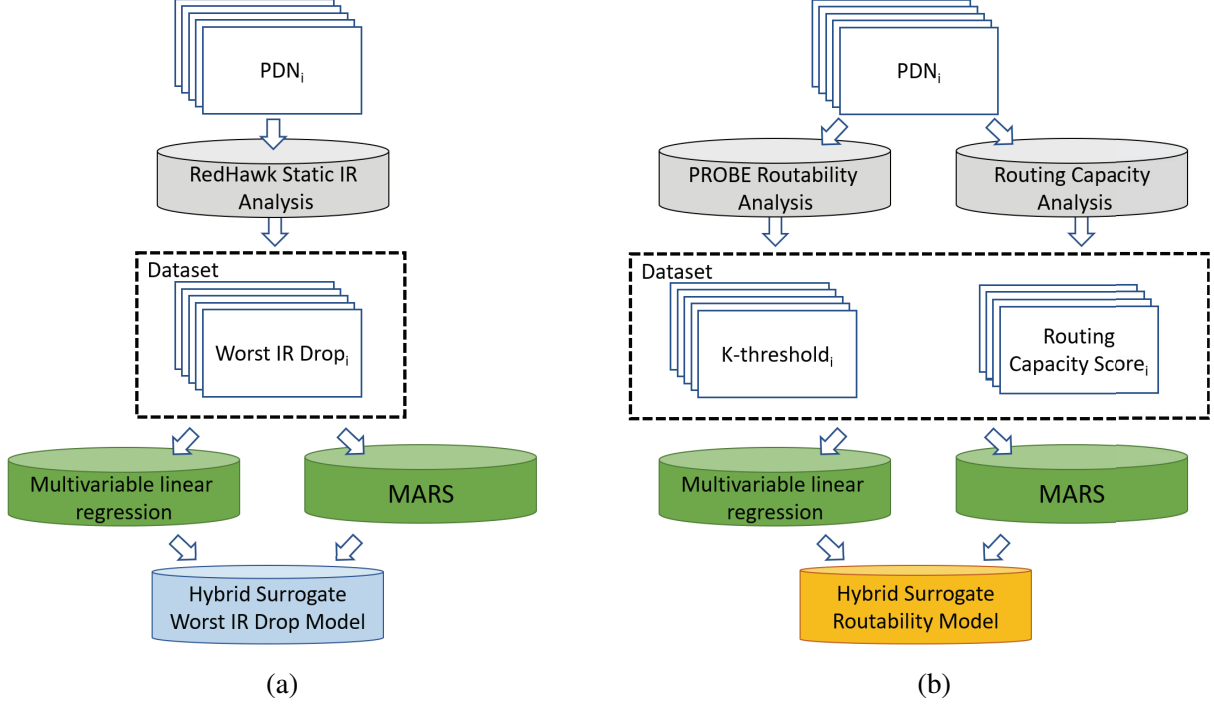


Figure 4.3: (a) WIR modeling flow. (b) Routability modeling flow.

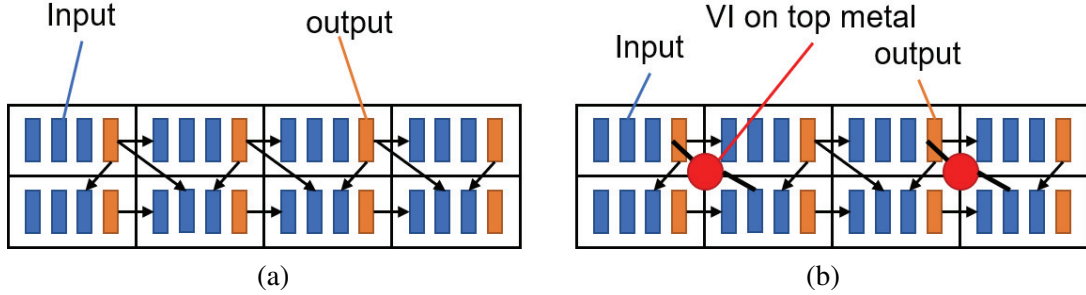


Figure 4.4: Illustration of (a) mesh-like placement as in [45], and (b) our 3D mesh-like placement with VIs.

We use both circuit design-independent knobs and circuit design-dependent knobs as inputs to our routability model. Furthermore, we consider the routing capacity scores [46] of all metal layers as additional inputs. We use the routability model to sort all PDN designs that satisfy the WIR requirement in order to find the optimal PDN.

In our routability modeling, it is important to rank the relative routability by the K_{th} value over the

absolute value of K_{th} predicted through regression. Therefore, not only the linearity expressed by the Pearson correlation coefficient [52] but also the ranking comparison by each K_{th} is required. We use the Spearman's rank correlation coefficient [55] to compare the routability ranking of PDNs with predicted K_{th} values through routability modeling and the ranking of PDNs with real K_{th} values obtained experimentally from PROBE-like routability analysis. The Spearman's correlation coefficient can be expressed using the following formula;

$$r_s = \frac{cov(rgX, rgY)}{\sigma_{rgX} \cdot \sigma_{rgY}}$$

where given a sample of size n , the n raw values X_i and Y_i are converted to ranks rgX_i and rgY_i , and $cov(rgX, rgY)$ denotes the covariance of the rank variables. σ_{rgX} and σ_{rgY} are the standard deviations of the rank variables. In general, there is a strong correlation when the coefficient between the two ranking groups is ≥ 0.9 .

4.3 Experiments

In this section, we describe our experimental setup and results. We perform experiments with a 28nm foundry technology library. We use 8-track cells in 28nm technology, and row utilization is determined by the number of available tracks. For example, utilization 0.727 is 8 tracks for cell and 3 tracks for white space.³ For PROBE-like routability study, we perform place-and-route using *Cadence Innovus Implementation System v18.10* [63]. For IR drop study, we perform static IR analysis using *ANSYS RedHawk v15.1.1* [64]. Table 4.2 shows the reference design we use for our experiments. For each model, we use 67% of the overall dataset for training and the remaining 33% of the dataset for testing. We use a MARS implementation in Python3 from the Py-earth package [57]. In this section, we show (i) scalability study; (ii) sensitivity study; (iii) IR drop model; (iv) routability model and (v) verification on real design.

Table 4.2: Reference design of PDN.

PDN design				
Metal layer	Direction	width (μm)	spacing (μm)	pitch (μm)
M2	H	Standard cell power rails		
M3	V	0.4	10	20
M4	H	0.4	0.8	12
B1 (M7)	V	8.0	16.0	60
B2 (M8)	H	10.0	20.0	70
Circuit design				
#Insts	25000			
Utilization	0.7			
VI density	0.05			

4.3.1 Scalability study

In this subsection, the scalability of our approach by varying design size is described. We perform routability analysis using variations of the reference PDN design. WIR in 3D IC depends on specific boundary conditions. We experimentally confirm that there is no obvious correlation between #inst and WIR for a given utilization. We sweep the number of cells from 25K to 100K with a step size of 25K for a fixed utilization with a total of 24 #PDNs with 75% (small) and 175% (big) design-independent knobs respectively. Our study results are shown in Figure 4.5. We observe that routability decreases as we increase the design size. Although there is a change in the absolute value of K_{th} when design

³For ease of use, the values of the following utilizations are rounded to the first decimal place.

size changes, the routability rank ordering of PDN designs remains the same. Based on the scalability observations, we fix the number of instances = 25K for the reference design in all experiments reported below.

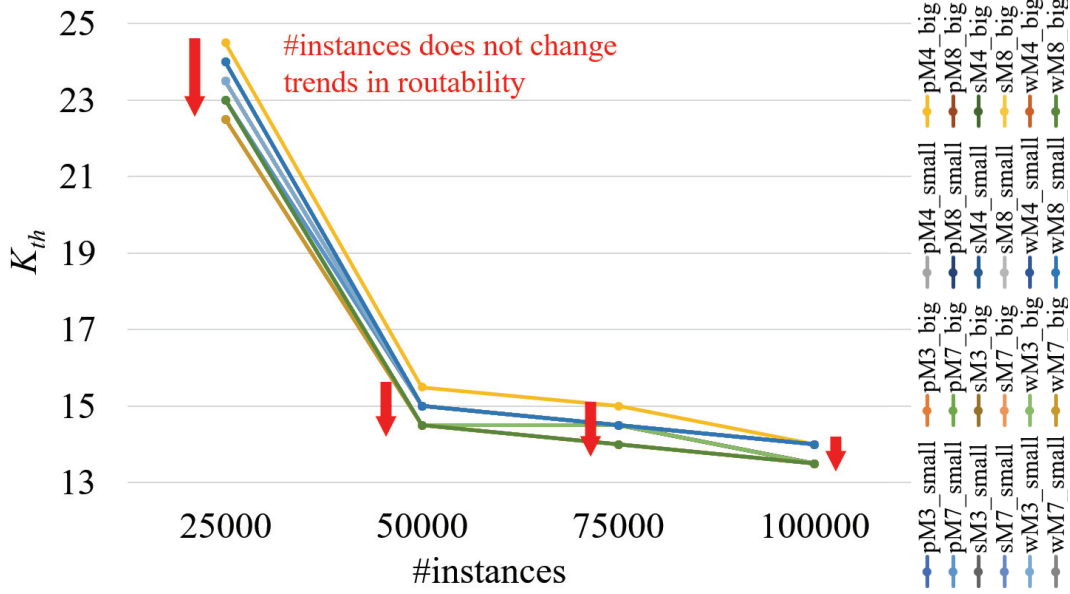


Figure 4.5: Routability (K_{th}) versus #inst reflecting various number of PDNs.

4.3.2 Sensitivity study

To assess the impact of each PDN and circuit design knob on WIR drop and routability, we investigate the sensitivities of worst IR drop and routability to various design knobs as discussed in Section 4.2.2. For PDN design knobs, all circuit-independent design knobs of *width*, *spacing* and *pitch* for M3, M4, M7 and M8 are considered. For circuit-dependent design knobs, we consider utilization and VI density. Only one knob is swept at a time while all other knobs are kept as shown in the reference design. Figure 4.6 shows the sensitivity results between WIR / routability (y-axis) and PDN density (x-axis) by varying design knobs. The PDN density of each layer is calculated by $2 \times \text{width} / \text{pitch}$

Width: We sweep *width* for M3, M4, M7 and M8 from 75% to 175% of the reference value. Figure 4.6(a) shows the worst IR drop as a function *width* for M4, M7 and M8 separately. We observe that worst IR drop decreases as we increase the *width* since VDD/VSS stripes are becoming less resistive. Figure 4.6(b) shows the routability as a function of *width*. For all layers, the routability of BEOL decreases as the *width* increases because less routing resource is available. Moreover, as shown in Figure 4.6(b), the higher layer has less sensitivity for the routability of the PDN layer utilization.

Spacing: We sweep the VDD/VSS stripes *spacing* for M3, M4, M7 and M8 from 75% to 175% of the reference value. Since the space between VDD and VSS stripes is mainly used to reduce dynamic IR drop and it does not have a significant effect on static IR drop. The effect of spacing on routability is also negligible.

Pitch: We sweep the M4 VDD/VSS stripe *pitch* for M3, M4, M7 and M8 from 75% to 175% of reference value. Figure 4.6(c) shows the worst IR drop as a function of *pitch*. We observe that worst IR drop increases as we increase *pitch* (i.e., sparser power mesh). Figure 4.6(d) shows the routability according to *pitch*. As with the case of *width*, routability decreases as PDN layer utilization increases. However, *pitch* has a higher sensitivity than the case of *width* even with the same PDN layer utilization.

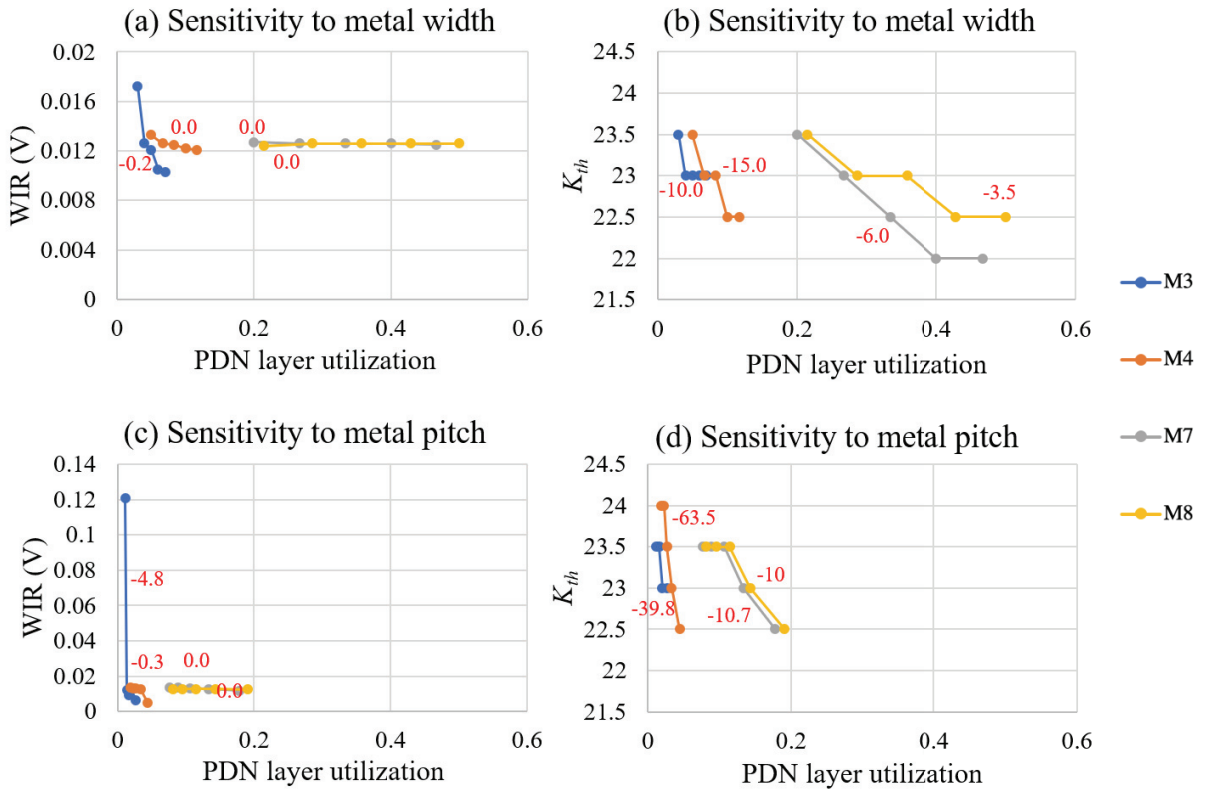


Figure 4.6: WIR (left) and routability (right) sensitivity results for circuit-independent knobs width (top) and set-to-set pitch (bottom). The red numbers indicate the slope of the K_{th} change with each knob.

Utilization: For our routability study, we use mesh-like placement. Therefore, current density is proportional to utilization in our study. Figure 4.7(a) shows the WIR as a function of utilization and metal width and Figure 4.7(c) shows WIR as a function of utilization and metal pitch, on M3, M4, M7 and M8. We observe that since IR drop is proportional to current density, which is in turn proportional to

utilization in a uniform placement, we have that WIR is proportional to utilization.

Designs with higher utilization tend to have DRVs on lower metal layers due to lack of routing resources for pin access and/or promotion. Therefore, we simultaneously sweep design utilization and metal width (resp. pitch) to study the routability of PDN design due to interaction between design utilization and width (resp. pitch). Figure 4.7(b) shows the routability as a function of utilization and metal width, and Figure 4.7(d) shows the routability as a function of utilization and metal pitch, on layers M3, M4, M7 and M8. We observe that routability decreases as we increase the utilization. We also observe that for the same utilization, routability is more sensitive to changes in lower metal layers.

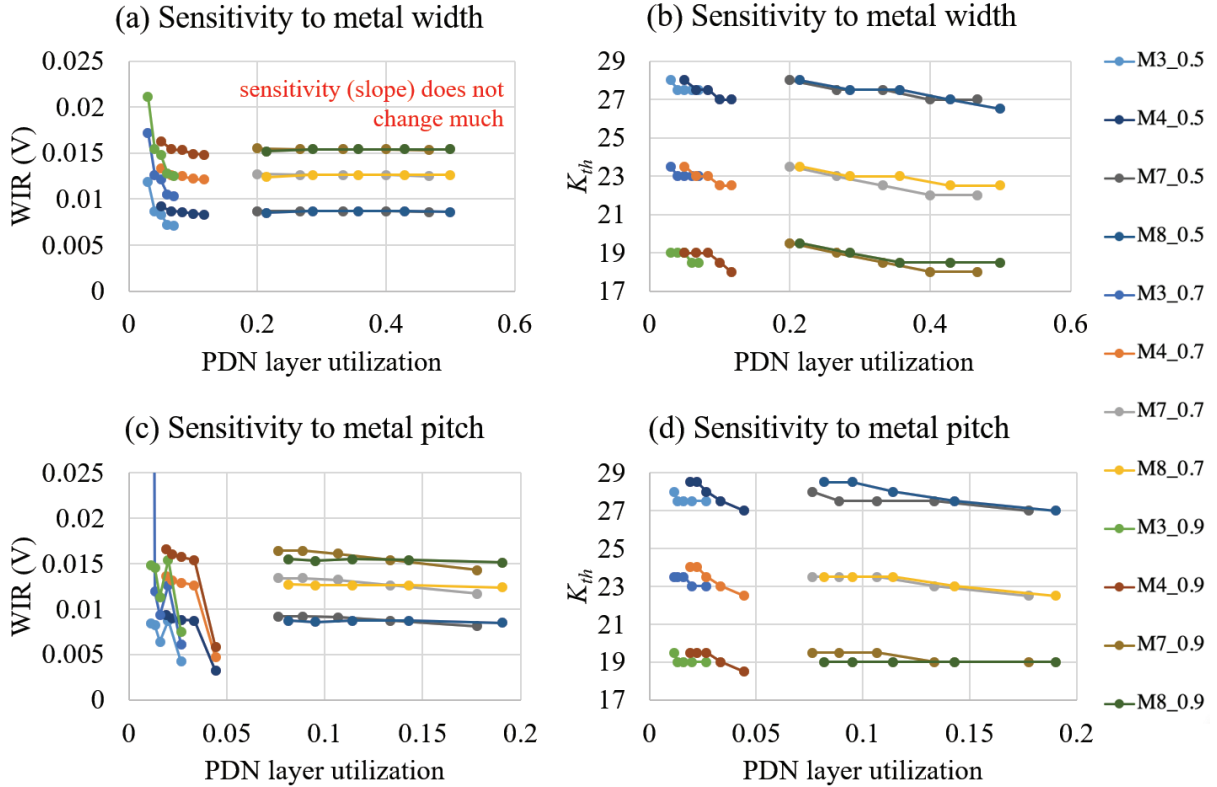


Figure 4.7: WIR (left) and routability (right) sensitivity analysis results for circuit-independent knobs width (top) and set-to-set pitch (bottom) with various utilization.

VI density: We sweep the VI density from 0.025 to 0.25. Similar to the utilization sensitivity study, we simultaneously sweep metal width or pitch along with VI density as VI accessibility, intuitively, depends more on routing resource on higher metal layer. Since signal VIs are circuit-dependent knobs that affect only routing resources, only the routability analysis is performed. (There is a slight difference between the target and actual VI density, because the VI should be aligned to the cell grid in mesh-like placement to guarantee the same distance between the VI and the connected net.) VI density is given in Table 4.3.

Figure 4.8(a) shows the routability as a function of VI density and metal width, and Figure 4.8(b) shows the routability as a function of VI density and metal pitch. We observe that routability suddenly decreases as we increase the VI density. Moreover, for a given VI density, routability is more sensitive to changes in higher metal layers, as we might expect.

Table 4.3: Sensitivity to VI densities (#nets = 25172).
#Nets = 25172

Target VI density	#VIs	VI density
0.025	684	0.027
0.05	1242	0.049
0.075	2052	0.082
0.01	2736	0.011
0.025	6266	0.025

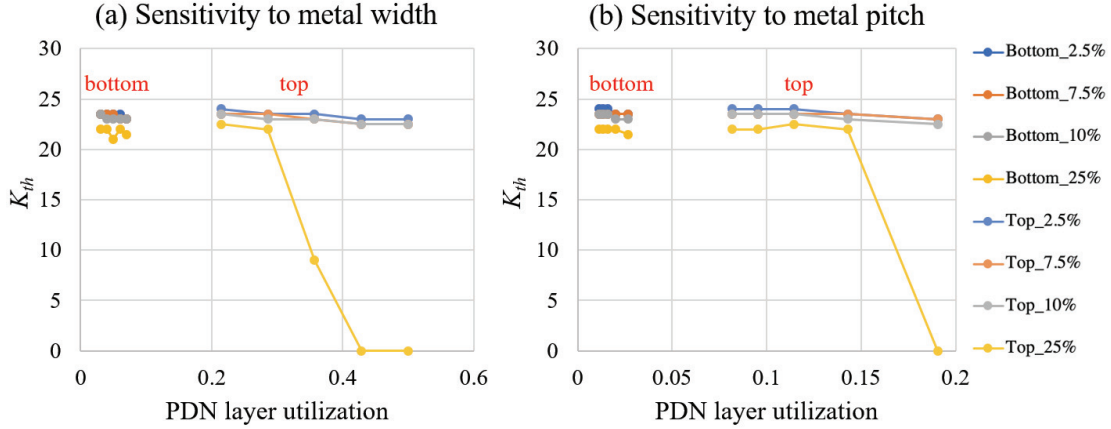


Figure 4.8: Routability sensitivity analysis results for circuit-independent knobs (a) width and (b) set-to-set pitch with various VI densities.

4.3.3 IR drop model

To efficiently assess whether a PDN design satisfies the worst IR drop requirement, we build an IR drop model based on a dataset which includes combination of knob values from *width*, *pitch* and utilization. Table 4.4 summarizes PDN and circuit values in the dataset. For the knobs that we explore in our experiment, we sweep the value of each knob from 75% to 125% of its reference value. Figure 4.9 shows the actual versus predicted WIR for various PDN designs with combinations of PDN design knob values. Our model achieves an absolute average error of 4.02mV (resp. 4.05mV) for the training (resp. testing) dataset.

Table 4.4: Summary of design solution space.

PDN design		
Metal layer	width (μm)	pitch (μm)
M2	Standard cell power rails	
M3	0.3, 0.5	15, 25
M4	0.3, 0.5	9, 12
B1 (M7)	6.0, 10.0	45, 75
B2 (M8)	7.5, 12.5	52.5, 87.5
Circuit design		
#Insts	25000	
Utilization	0.7	
VI density	0.05	

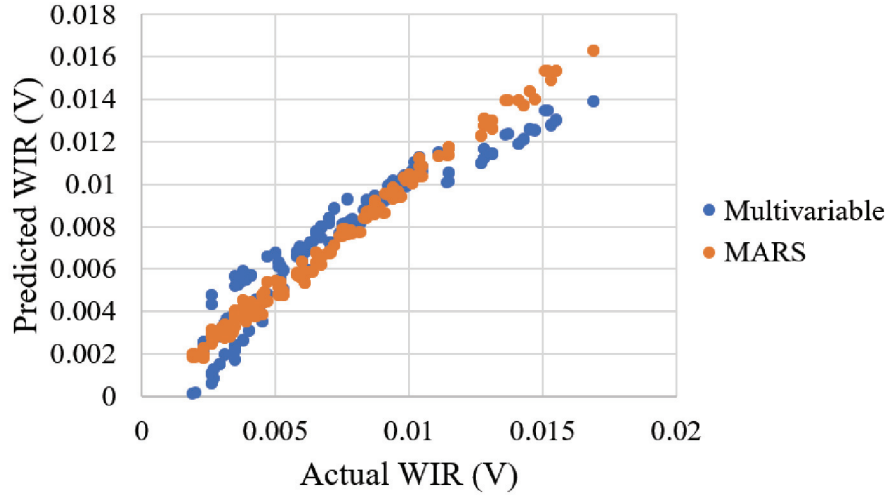


Figure 4.9: Results for WIR model.

4.3.4 Routability model

To find the optimal PDN pathfinding, we rank PDN designs that satisfy worst IR drop requirement by routability. We use the same dataset in Section 4.3.3 to build a routability model. We consider the following three models in this work.

- **Model1:** knob-based model
- **Model2:** routing capacity score-based model
- **Model3:** combined model

We build **Model1** based on the knob values of a circuit design for a given PDN. The input of the

model is a sequence of PDN design knobs for all metal layer in BEOL stack followed by circuit design knobs. Second, we build **Model2** based on routing capacity score of each metal layer [46]. Third, we build **Model3** based on both knob values and routing capacity scores together.

Figure 4.10 illustrates correlation between actual K_{th} and predicted K_{th} by each knob-based model. As shown in Figure 4.10, accuracy of predicted routability results by **Model2** is not as good as that of **Model1** or **Model3**, because the routing capacity score calculated based on the *gcell* grid does not consider the capacity difference due to the variation of PDN width and pitch of each layer. On the other hand, training and testing within the same knob sweep range cannot show the generality of the model. Thus, we cannot directly compare **Model1** and **Model3** using the results in Figure 4.10.

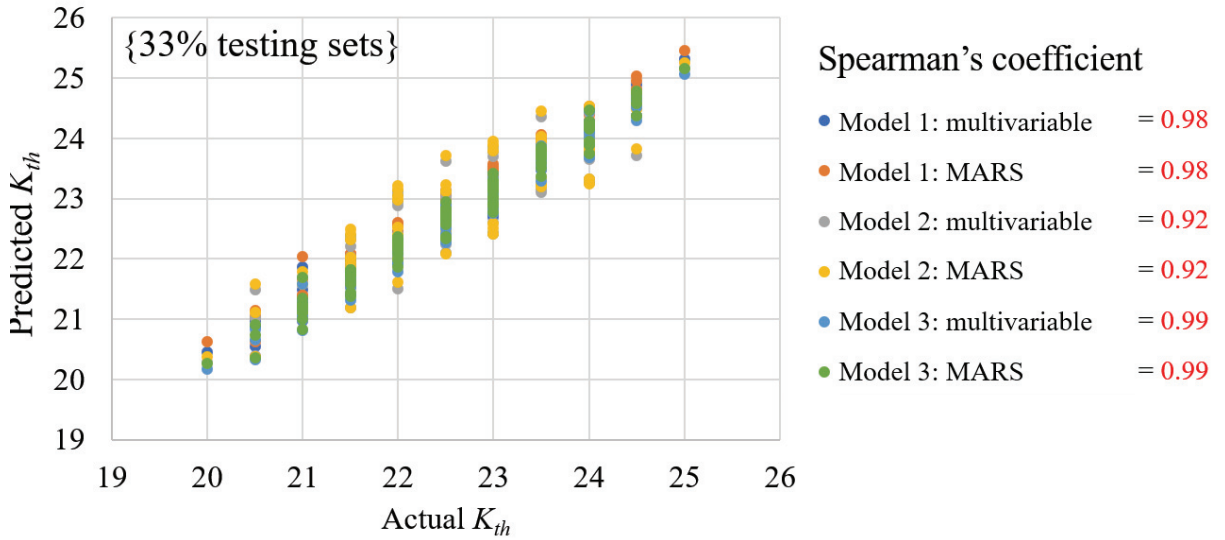


Figure 4.10: Correlation of routability graph between actual K_{th} and predicted K_{th} by each knob-based (**Model1**), routing capacity score-based (**Model2**) and combined (**Model3**) models. The scatter points displayed in the graph represent a total of 88 #testing points and a total of 168 #PDNs training points.

In order to compare and verify the generality of the model, we build another routability model based on a dataset that is composed of routability data with knob values of 85% and 115% of its reference value (i.e., a “subset” of the dataset in Table 4.4). We then apply the model built from the “subset” on the Table 4.4 dataset. Figure 4.11 shows the routability modeling results. Our models built from a “subset” dataset have reasonable accuracy on the original dataset, which implies that our model can be generalized and used for other testcases with an interpolation and extrapolation. Figures 4.11(a) and (b) show the case of **Model1**. We achieve a Spearman's coefficient of 0.96 and 0.91 with multivariable linear regression for testing dataset. This is a strong correlation between the two ranking groups. Fig-

ures 4.11(c) and (d) illustrate **Model3**, but compared to **Model1**, the accuracy is lower. For an accurate model through regression analysis, all x-axis variables should be independent of each other. However, since the variables of **Model2** depend on the PDN design knob, the combined **Model3** is worse than **Model1**.

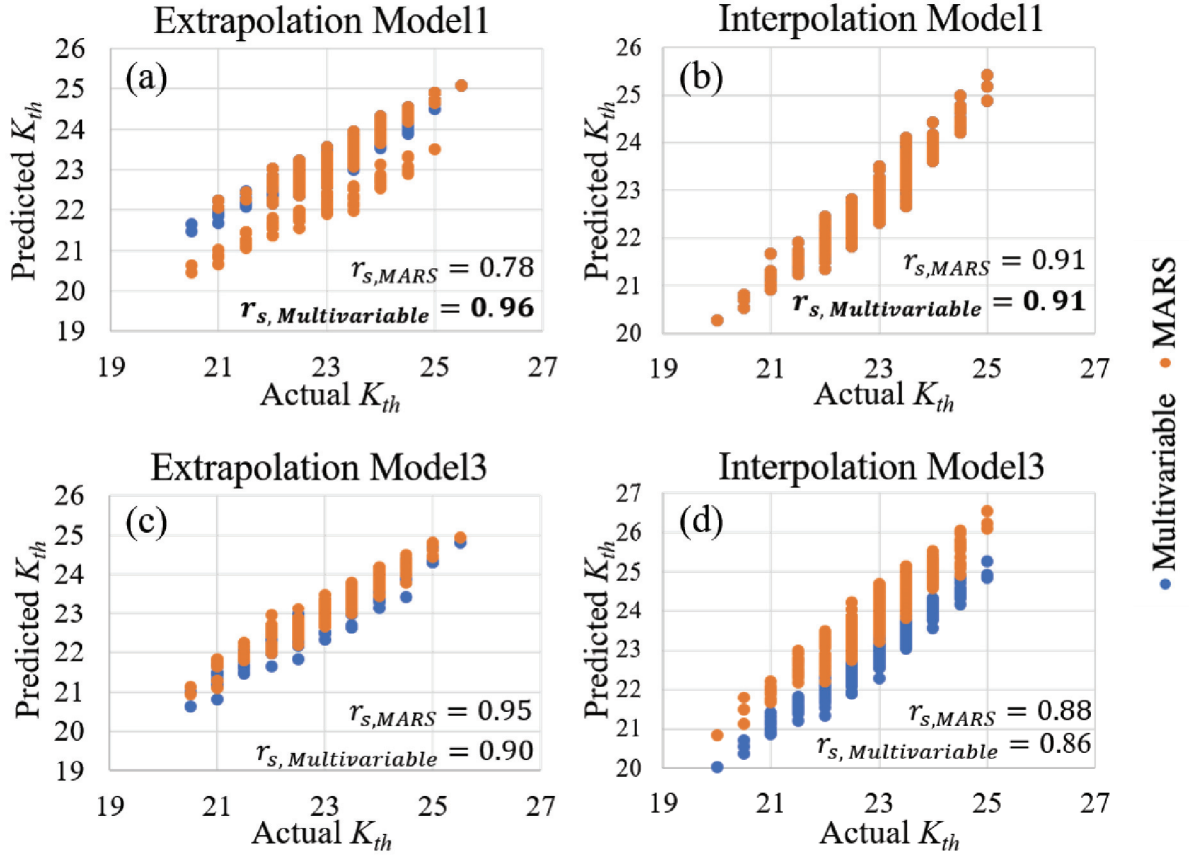


Figure 4.11: Correlation of routability graph between actual K_{th} and predicted K_{th} values by each of (a), (b) knob-based (**Model1**) and (c), (d) combined (**Model3**). The scatter points displayed in the graph represent a total of 256 #testing points and a total of 256 #PDNs training points.

4.3.5 Verification of pathfinding on real design

We verify the routing capacity and WIR drop model by applying pathfinding methodology to a real design testcase. We use AES encryption core and JPEG encoder from OpenCores [62], and information of testcase is described in Table 4.5. Each design is synthesized with *Synopsys Design Compiler L-2016.03-SP4-1* [65]. We perform experiments with 8-track standard cells from a 28nm foundry technology library. Since each cell of real designs does not have uniform width unlike mesh-like placement,

we perform legalization before routing to eliminate overlap caused by random neighboring cell swapping. We set the fixed utilization as 0.727 because the scalability of the model was already shown in Section 4.3. To use the proposed routability model, we add VIs as I/O pin then the pins are uniformly placed on the top metal equal to the VI density used in the model (5% of #VIs/#nets). The additional VIs are connected to the nearest different nets. Without loss of generality, we use the WIR value of reference PDN design as IR drop requirement for each testcase. The BEOL stack of the PDN is the same as the reference PDN of Table 4.2.

Table 4.5: Information of testcases.

Testcases	Clk period	#Insts	#Nets	Util	#VIs (~5%)
AES cipher top	1.4 ns	10623	10882	0.727	529
JPEG encoder	1.4 ns	23884	27161	0.727	1292

Based on the trained routability model, PDNs with an IR drop greater than IR drop for reference PDN are filtered, then the design knobs that constitute the best PDN can be obtained through the predictive model. To verify the ranking of the routability model, we pick two best PDNs, two intermediate quality PDNs, and two worst quality PDNs for verification with real design. Table 4.6 shows verification results with AES cipher and JPEG encoder testcases. In actual designs, the cell placement is not uniform, so the denoising is performed through five different random seeds, and the K_{th} of Table 4.6 is the average value of five runs. As shown in Figure 4.12, although the predicted Kth value and the actual Kth absolute value are different depending on the netlist, the ranking order is maintained. Thus, an optimal PDN with design knobs is found that has an IR drop less than the reference PDN, along with a best routability.

Table 4.6: Simulation results with real testcases for AES cipher and JPEG encoder. All K_{th} values are average values with five de-noising runs.

PDN	AES				JPEG			
	clk (ns)	#inst	K_{th}	IR drop (V)	clk (ns)	#inst	K_{th}	IR drop (V)
Best	1.4	10k	2.08	0.0113	1.4	24k	13.38	0.0431
Reference			1.90	0.0129			11.90	0.0438
Worst			1.76	0.0078			9.08	0.0309

4.4 Conclusions and Future Directions

In this work, we present a novel power delivery pathfinding methodology for emerging die-to-wafer face-to-face integration. Our work offers several advances as compared to previous works: (i) BEOL routability analysis with consideration of pre-routed PDN; (ii) a new study of interactions between IR drop analysis and routability analysis; and (iii) a PDN pathfinding flow that identifies a high-routability, satisfying PDN design with respect to prescribed worst IR drop constraints for a given design and given BEOL stack options. Experimental studies confirm the stability of the routability ranking of PDN design options across design sizes, as well as “scale-independence” of IR drop behavior due to regular power and ground TSVs; these phenomena are enabling to our pathfinding strategy. In experiments with a 28nm FDSOI enablement, the pathfinding model also accurately predicts the most routable PDN satisfying prescribed IR drop limits. Our ongoing and future works include (i) exploration of additional BEOL stack options, particularly in advanced nodes, and (ii) extension of our approach to integration technologies other than face-to-face integration.

4.5 Acknowledgments

Chapter IV is in part a reprint of “Power Delivery Pathfinding for Emerging Die-to-Wafer Integration Technology”, *Proc. IEEE/ACM Design, Automation and Test, in Europe* 2019.

I would like to thank my coauthors Professor Andrew B. Kahng, Professor Seokhyeong Kang, Dr. Kambiz Samadi and Bangqi Xu.

I also thank Dr. Hyein Lee for helps on the benchmark generation.

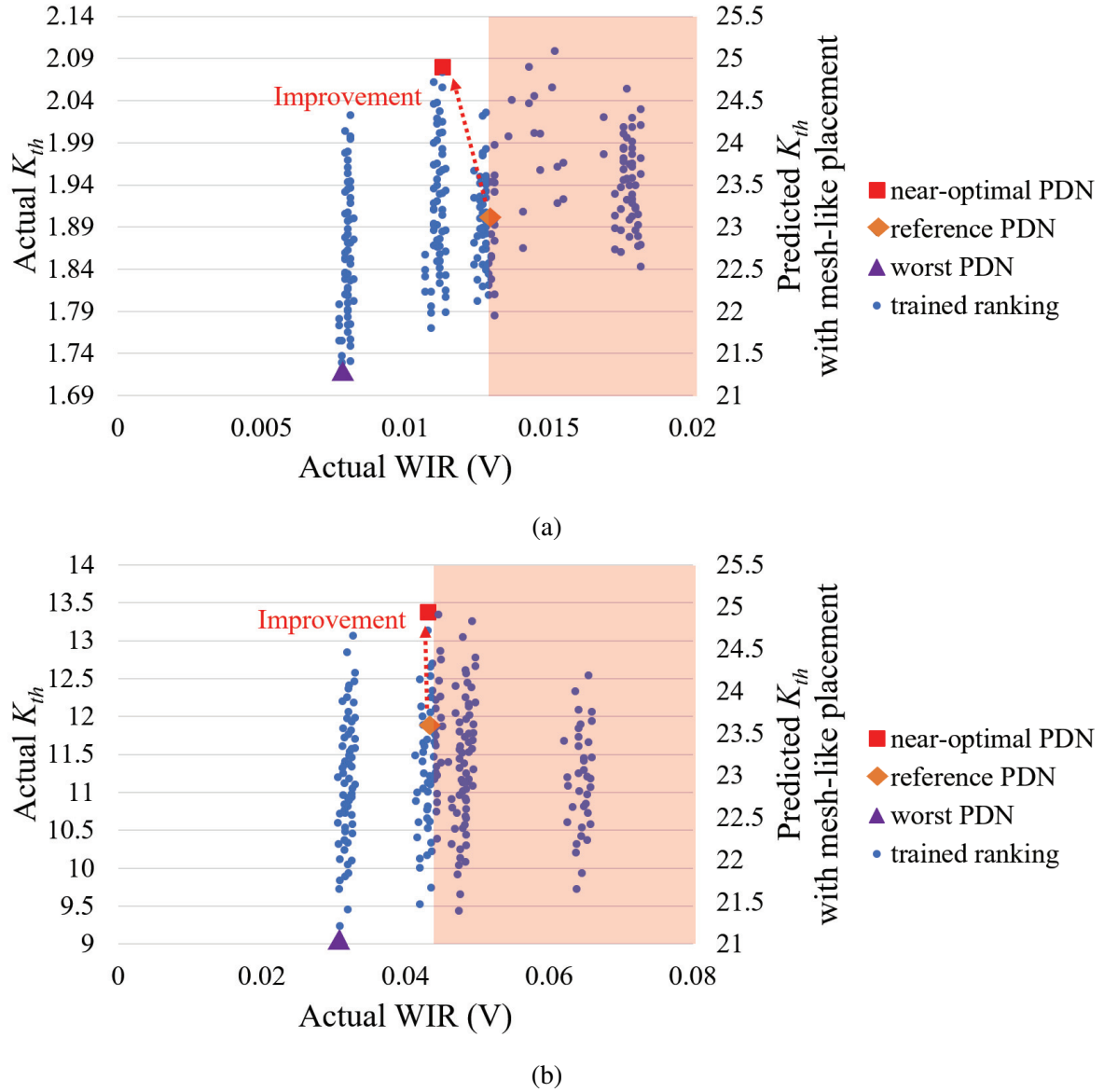


Figure 4.12: Routability (K_{th}) versus IR drop data with PDN design knobs for (a) AES encryption core and (b) JPEG encoder testcases. Blue dots denote trained ranking of PDNs and are represented by the second y-axis as K_{th} values. Optimal, reference and worst PDNs are verified by real designs. The red arrows indicate improvement from the reference PDN. The red region indicates WIR greater than the WIR drop of the reference PDN.

Chapter V

Power Integrity Coanalysis Methodology for Multi-Domain High-Speed Memory Systems

Main interests on the design of PDNs have been in the simultaneous switching noise (SSN) caused by the power plane resonances and the voltage (IR) drop due to the resistive effect [68]. However, in the upcoming mobile design environment, the coupling between separate power domains should be considered as well [69]. We present an initial study on the power domain coupling, focusing on the effect of grounds and the substrate design parameters. For a simple package PDN structure, the effect of geometry on the coupling can be clarified by observing equivalent capacitances, which are extracted from the impedance data when frequency is less than 1 GHz.

In this chapter, we propose a PI design methodology that considers the effects of the electrical and structural parameters on the multi-voltage domain chip-package-PCB system as shown in Figure 5.1. For PI analysis, we construct a parametric link model that can quickly and accurately predict electrical performances according to the various design variables of a high-speed link. In addition, we develop a statistical analysis of the performance data derived from the link model that provides design guidelines. Thus, we propose a parametric link simulation environment that considers the electrical characteristics of the on-chip high-frequency switching current, and the physical effects of the package-PCB in multiple power domains. We verify our proposed coanalysis methodology by using the *SPICE* results under the Joint Electron Device Engineering Council (JEDEC) LPDDR4 environment with an industrial results. We derive the results of the transient and AC simulation by using the proposed model based on the

parameters that reflect the characteristics of the power delivery system (PDS). The main contributions of this chapter are as follows:

- We propose a coanalysis methodology for PI, and verify it by using a full layout *SPICE* simulation under the JEDEC LPDDR4 environment.
- We also propose a pseudo-random current profile that characterizes realistic on-chip current profiles with that the desired design constraints.
- We analyze the effect of (1) the domain coupling in the package, (2) the on-chip decoupling capacitor, and (3) the input noise for power integrity with sweep or Monte Carlo simulations of the design parameters.
- Our proposed PI coanalysis methodology can perform fast and massive simulations based on the various design parameters of the PDS.

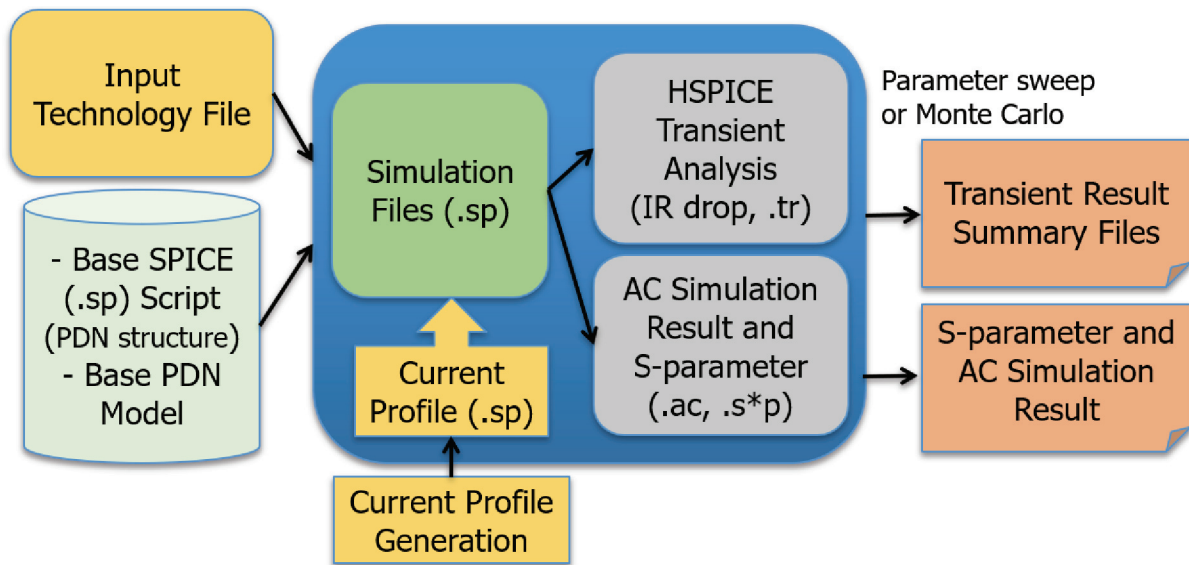


Figure 5.1: Process of our proposed PI coanalysis methodology.

5.1 Power Delivery System Analysis Model

A common method for constructing a memory system is the traditional discrete isolated-chip packaging for disposing a SoC and memory in two dimensions. However, in a mobile system that requires low power and high integration density, the package-on-package (PoP) stacking method with short vertical interconnection is the preferred application [75, 76]. Therefore, we implement the PoP method to model a PDS between the SoC and memory chip as shown in Figure 5.2.

In the PDS model, the power is transferred from the voltage regulator module (VRM) to the memory chip through a PCB board, an SoC package, and a memory package. The PCB and each package include decoupling capacitors to reduce the frequency-dependent noise; they also include a multi-domain PDN model. We assumed wire bonding between the package and the chip. We modeled the on-chip with the switching current by using the memory operation as the redistribution layer (RDL) and the current profile. The following subsections describe the details of each component of the PDS model.

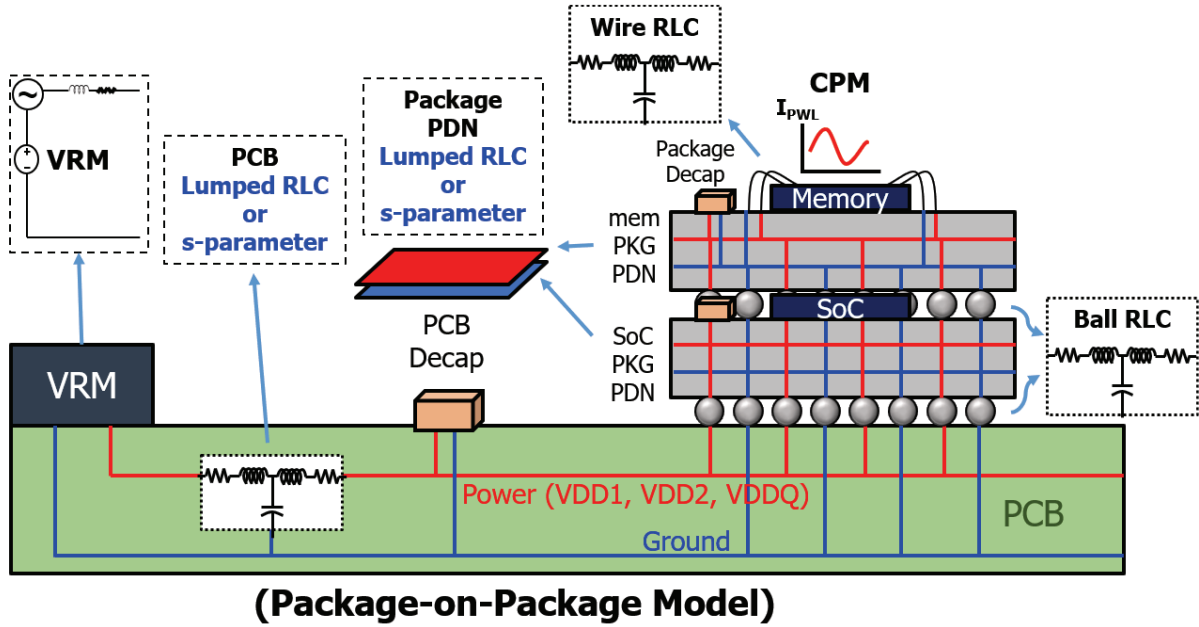


Figure 5.2: Overview of the PDS for a PoP model.

5.1.1 Multi-domain power distribution network

There are various methods for modeling a package-level PDN in a general single-power domain based on the impedance calculation [77] and the extraction from the physical structure and application to the equivalent resistor-inductor-capacitor (hereinafter RLC) model [80, 78, 79]. To analyze how the

changes of the PDN structure affect the power domain coupling, we use a modeling method that extracts the S -parameters based on the physical structure. To prevent coupling between the power domains, a ground plane is placed between them in general. We use a PDN structure with two power domains as shown in Figure 5.3. This PDN structure has two power domains and ground planes. The two power domains are surrounded by a ring ground and placed separately next to the center ground plane. Each ground plane is connected to the bottom ground plane by vias. The design parameters are described in Figure 5.3 and Table 5.1. This PDN structure is electrically small and simple compared with the realistic PDN geometry, and the location of the port in the same power plane is not dominant. Therefore, we choose the input/output ports of the two domains, as shown in Figure 5.3 (left). We mainly consider three design parameters to analyze the domain coupling effect. First, we change the existence of the central ground (hereinafter CGND) and the ring ground (hereinafter RGND). Second, we vary the width of the edges around the power domain planes (MARGIN_WIDTH); the other dimensions are keep fixed.

5.1.2 Input power source

The VRM consists of a DC-DC converter and a feedback control circuit that supplies the reference voltage required by the system for the output stage. The VRM is basically a nonlinear system. However, when the nonlinear model is implemented in the power transfer model, the simulation runtime becomes long, and it becomes difficult to set the parameters that determine the characteristics of each element in the VRM. Therefore, when the power transfer model is constructed to analyze the power integrity, the VRM is simplified into a linear model. The buck switching regulator is a widely used nonlinear VRM model [82], which we simplify into a linear model for a *SPICE* simulation. For the VRM, a four-element linear model can be used, as shown in Figure 5.4(a). However, the four-element model is a case-dependent model, and L_{slew} and R_{flat} need to be extracted from the nonlinear model; therefore, it is difficult to determine the value of the elements at the stage in which the PCB impedance has not yet been determined [83]. Our aim is to analyze the PDS model and treat the VRM as an input source. We use simplified two-element models instead of predicting the PCB impedance in advance, as shown in Figure 5.4(b).

Based on the two-element linear model, we add the sinusoidal noise source and the VDD offset. L_{out} is the inductance value of the cable between the VRM and the system board; it affects the maximum effective frequency. R_0 represents the resistance between the VRM sense point and the actual load. The R_0 of VRM can be calculated as follows:

$$R_0 = \rho \frac{l}{A} \quad (5.1)$$

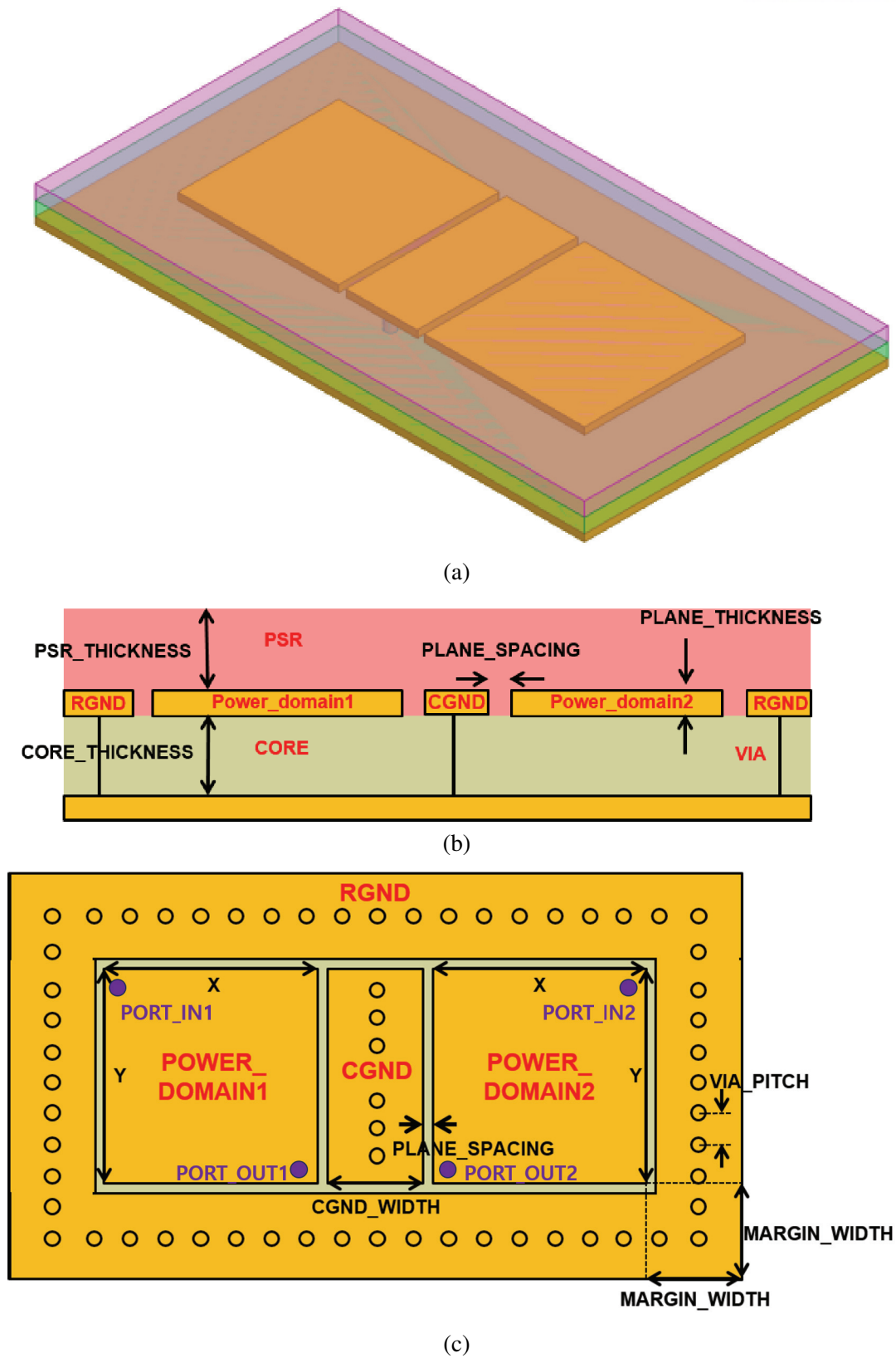


Figure 5.3: Structure and parameters of a sample PDN: (a) slant view, (b) side view, and (c) top view.

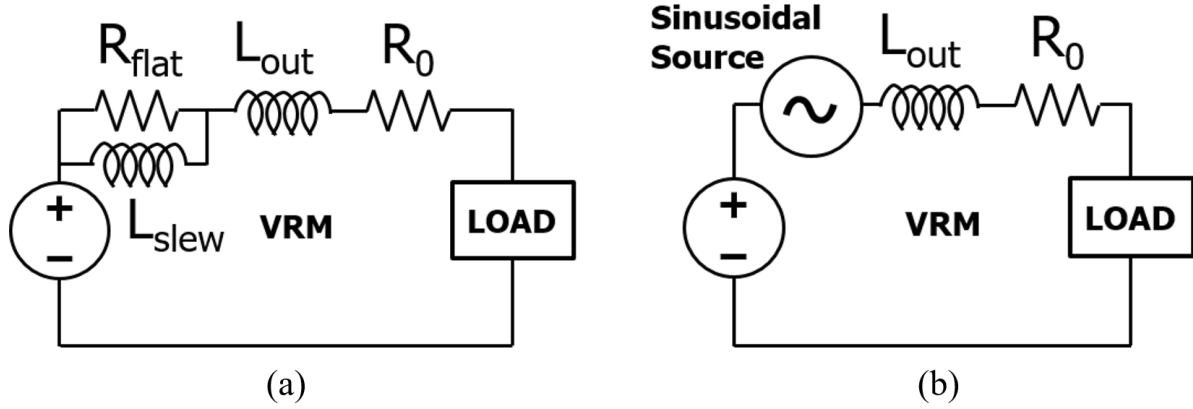


Figure 5.4: (a) Four-element linear VRM model and (b) the proposed two-element linear input power source model with a sinusoidal source as the noise.

where ρ is the resistivity (assuming copper is used in this model); A is the cross-sectional area of the die; and l is the maximum path length.

We model the input noise itself as a sinusoidal source to make it easier to use in EDA analysis. The noise ripple and the VRM frequency is set in the sinusoidal source, and the DC voltage level is determined by setting the VDD offset. In this way, the PDS input can be adjusted as an environment similar to the VDD source output from the actual VRM. The sinusoidal source can be formulated as follows:

$$V_0 + V_a \cdot \sin \left[2\pi \left[Freq \cdot (time - T_d) + \frac{Phase}{2\pi} \right] \right] \cdot e^{-(time - T_d) \cdot D_f} \quad (5.2)$$

The parameters are shown in Table 5.2.

5.1.3 On-chip modeling

On-chip redistribution layer

The On-chip RDL is used to route the wire from the bonding pads to the bump pads without changing the position of the I/O pads. The RDL is placed on the top metal layer of the die. In our PDS analysis model, we use a simple lumped RLC T-model because the capacitance and inductance of the RDL are relatively small in the overall system (see Figure 5.5). Note that the lumped RLC value is obtained from the real RDL layout, and the resistance value significantly affects the static IR drop of the VDD as compared with the capacitance and inductance. The T-model is electrically bi-directional, which is closer to the realistic PDN model and has higher accuracy than the RLC model in AC analysis [84]. The RDL model is connected to the memory package wire and the on-chip pad model that consists of a

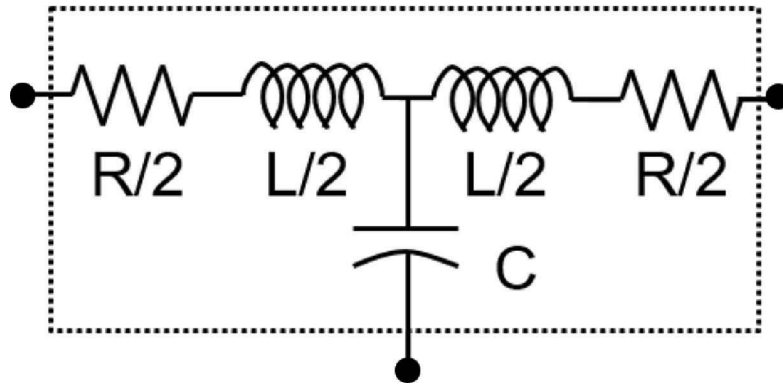


Figure 5.5: Lumped RLC T-model.

lumped resistor and a capacitor.

We use a pseudo-random current profile that reflected the behavior of the real current in the chip's characteristics because it is difficult to implement a realistic signal channel or use an industrial current profile.

On-chip current profile generation

In general, to reduce the problems faced during simulation and runtime, PI analysis is performed by *SPICE* simulation using the chip power model (CPM) [85] with the input/output buffer information specification (IBIS) model [86] in the PDS instead of the realistic channel model [87]. The CPM contains information about the on-chip, and it is easy to use in a package and with the PCB design. However, it is difficult to coanalyze between the on-chip elements and the package design elements by using the CPM. For instance, the CPM already includes the RDL, the on-chip capacitance, and the metal resistance and inductance. Therefore, it has less flexibility in design simulation from the perspective of the package and the PCB designer.

On the other hand, a current profile is generated by referring to the target impedance of the PDS. Kim et al. [88] assumed that the maximum IC switching current of memory had a simple triangular shape for the peak current. However, this is an optimistic method because it is difficult to predict the PI due to parasitic capacitance and the inductance at high frequency in real PDSs. Chen et al. [89] implemented a memory controller hub and dynamic random-access memory modules to generate a current profile. However, creating a current profile through a memory module design is also difficult for the same reason as that for CPM. In addition, the current profile of the commercial memory is not available to the public. Therefore, in our methodology, we propose a pseudo-random current profile generation.

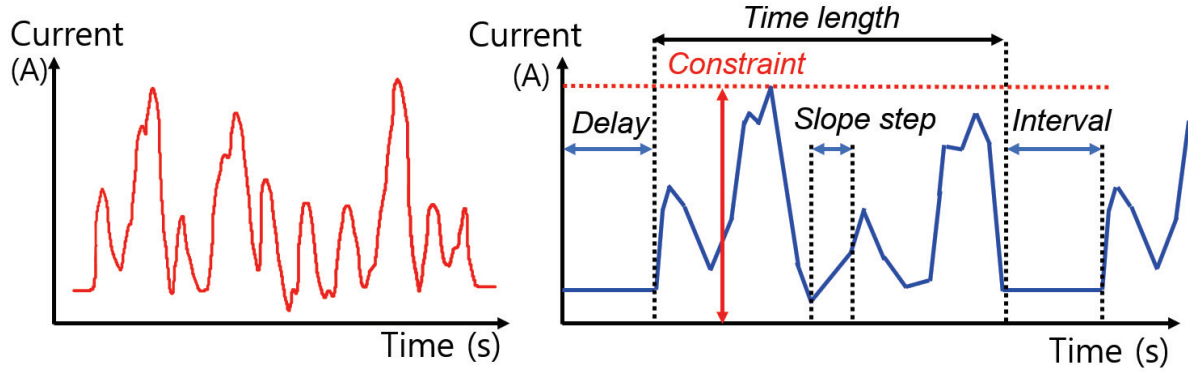


Figure 5.6: (a) Real on-chip current profile on the pad of VDDQ. (b) Overview of the characterized pseudo-random current profile parameters.

We first define the parameters that characterized the real on-chip current profile. Figure 5.6(a) shows the real on-chip current profile of VDDQ measured on the pad of the memory die. The characterized parameters to simulate the realistic operation are shown in Figure 5.6(b). We set the parameters so that they had a strong influence on the power transfer characteristics during realistic operations to achieve a current demand similar to a real memory operating environment. We take the min/max current constraint and the min/max *slope_step* value as the input parameters. In addition, we generate a randomized piecewise linear waveform to satisfy the input condition. The *slope_step* determines the time between the changes in the current slope (di/dt) as one step. A random current value between the min/max *current constraint* is arbitrarily set for each *slope_step*. Then, the waveform is generated by increasing or decreasing based on the random current value while satisfying constraint. The *delay* parameter is the initial delay time in which the current profile starts. The *interval* represents the time taken to repeat the current generated for memory operations, and the *time_length* is the period of the repeated current profile excluding the interval. Note that the above parameters are not intended to exactly imitate the realistic waveforms; however, they are intended to facilitate analysis through parametric simulation by characterizing factors that affect the PI.

5.1.4 Decoupling capacitor, wire and ball modeling

A real decoupling capacitor has equivalent series inductance (ESL) and equivalent series resistance (ESR). As shown in Figure 4.3, the decoupling capacitors are connected to the PCB, the SoC package, and the memory package. In addition, we connect the decoupling capacitors with ESL and ESR.

We connect the wires and balls to the VRM, the package, the PCB, and the load by using the simple RLC model as shown in Figure 5.5. This RLC model is applicable when the length of the interconnect

is electrically short, for example, in a ball structure. When the length of the wire exceeds 1/10th of the wavelength at the maximum modeling frequency (i.e., the knee frequency), the accuracy of the first-stage RLC model decreases. Therefore, we need to design the RLC model as a multistage series RLC network.

In the RLC model of the wire and ball, the value of the AC resistance (R) is calculated using the following equation [90]:

$$R \approx \frac{L\rho}{\pi(D-\delta)\delta}, \quad D \gg \delta, \quad (5.3)$$

where δ is the skin depth, which is calculated as

$$\delta = \sqrt{\frac{2\rho}{\omega\mu_r\mu_0}}, \quad (5.4)$$

where ρ represents the resistivity of the conductor; ω is the angular frequency; μ_r is the relative magnetic permeability of the conductor; and μ_0 is the permeability of the free space. This equation is an approximation that can be used in a high-frequency range where the skin depth is very small. If the frequency is low, the DC resistance must be considered together and the following equivalent resistor equation can be used:

$$R = \sqrt{R_{DC}^2 + R_{AC}^2} \quad (5.5)$$

The modeling frequency is determined based on the relative sizes of R_{DC} and R_{AC} , which are related to the cross-sectional area and the conductivity of the wire or the ball.

The self-partial inductance of a wire with the radius r_w and the length l is calculated as follows [91]:

$$L_p = 2 \times 10^{-7} l \left[\ln \left[\left(\frac{l}{r_w} \right) + \sqrt{\left(\frac{l}{r_w} \right)^2 + 1} \right] - \sqrt{1 + \left(\frac{r_w}{l} \right)^2} + \frac{r_w}{l} \right] \quad (5.6)$$

The self-inductance of the wire structure has frequency dependence because of the decrease in the internal inductance. However, when the radius of the wire is very small in comparison with the length, and the dielectric constant of the wire is also small, the changes in the self-inductance due to the frequency can almost be neglected.

The capacitance (C) was calculated using the following equation [92]:

$$C = \frac{2\pi\epsilon l}{\arccos\left(\frac{d}{a}\right)}, \quad (5.7)$$

where a is the wire radius; d is the distance; l is the wire length; and ϵ is the permittivity. This capacitance equation is established between a conductor with a circular cross-sectional area and the ideal ground. It cannot be generally applied when another conductor exists nearby, but it can be used approximately when the ground capacitance is dominant. If the dielectric constant is preserved according to the frequency, the capacitance equation is independent of the frequency. In addition, if the loss factor is considered in the permittivity of the PCB or the packaging medium, the conductance term also needs to be considered.

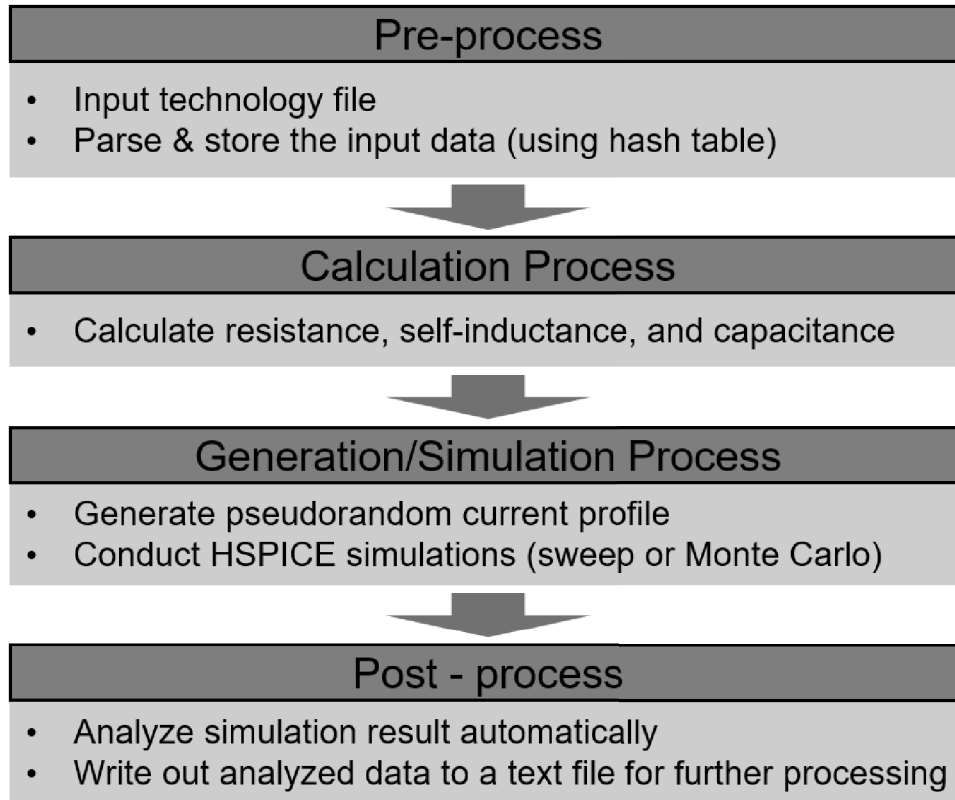


Figure 5.7: Procedure of our proposed fast analysis methodology for PI.

5.2 Procedure of Proposed Methodology

This section describes in detail the PI coanalysis methodology. As shown in Figure 5.7, we first read the PDN and simulation parameters and internally computes the RLC values in pre-calculations process. The PDN model is simulated using the *S*-parameter. If we directly input the lumped RLC value of the PDN, the generated lumped RLC PDN model and the remaining parameters are parsed. In addition, an on-chip current profile is generated based on the input parameters. All the simulation results are automatically summarized in the output file.

In our PDS model, PI can be analyzed using two methods that change the design and process parameters: sweep simulation and Monte Carlo simulation. In sweep simulation, we observe how each parameter affects the PI characteristics. The variation in the IR drop of the on-chip can be observed by sweeping one parameter. This allows us to analyze the relative impact of the different parameters and the trends of the linear and nonlinear effects. Monte Carlo simulation is mainly performed to investigate how the errors in the process affect the entire PDS. For example, when the geometric dimension of the memory package PDN varies, Monte Carlo simulation can be used for PI analysis. In addition,

our proposed methodology enables fast statistical analysis of simultaneous process/voltage/temperature changes through massive Monte Carlo simulations. Both the sweep and Monte Carlo methods are applicable for transient and AC simulations.

5.3 Simulation and Analysis

Our proposed methodology is written in *Perl* and *Synopsys HSPICE* [98], and the PDN model is extracted from the *HFSS* tool [99]. We validate using a 2.4 GHz Intel Xeon E5-2620V3 Linux workstation with a single core for the *SPICE* simulation. We compare our proposed methodology with the full layout *SPICE* transient simulation, and perform several case-based sweeps and Monte Carlo simulations for PI analysis. Figure 5.8 shows the simplified schematic of VDDQ and VDD2 of the PoP PDS model. Our multi-domain PDN model is adopted for the SoC and the memory package. We use PDN of the PCB board as the lumped RLC element with actual parameter values. All the simulations are measured at the pad of the on-chip to analyze the power transfer from the VRM to the memory chip (see Figure 5.8). The nominal value of the important parameters and the VDD (VDDQ/VDD2) noise values are described in Table 5.3. Other parameters have less effect on the results; therefore, we set the other parameters based on the 800 MHz LPDDR4 memory environment.

5.3.1 Preliminary analysis of power domain coupling

We first investigate our proposed multiple power domain PDN. Simulation frequency range is from 10 MHz to 1 GHz. Twenty ports are assigned based on the pad pitch on each power domain. Ten ports (from port 1 to port 10) are assigned in power domain 1, and the other ten ports (from port 11 to port 20) are assigned in power domain 2, as shown in Figure 5.9.

Figure 5.10 shows the impedance (Z -) parameters of a PDN design case. Since the structure is electrically small for the frequency range of our interest, the characteristics of ports on the same power domain are not distinguishable, as shown in Figure 5.10. Thus the Z -parameters can be classified to the self impedances and the mutual impedances. The self-impedance represents the capacitance of each power plane itself, whereas the mutual impedance results from the coupling effect between planes. The overall impedance characteristics also indicate that capacitance is dominant when considering an equivalent circuit of the coupled PDN structures. Hence, we can easily analyze the capacitance of 2-port network system by selecting $Z(1,1)$ for the self-impedance and $Z(1,11)$ for the coupling impedance between power domains.

Equivalent capacitance calculation

Since the network system is reciprocal, a 2-port network system of PDN can be constructed by self and mutual capacitors: C_{11} , C_{12} , C_{22} , as shown in Figure 5.11.

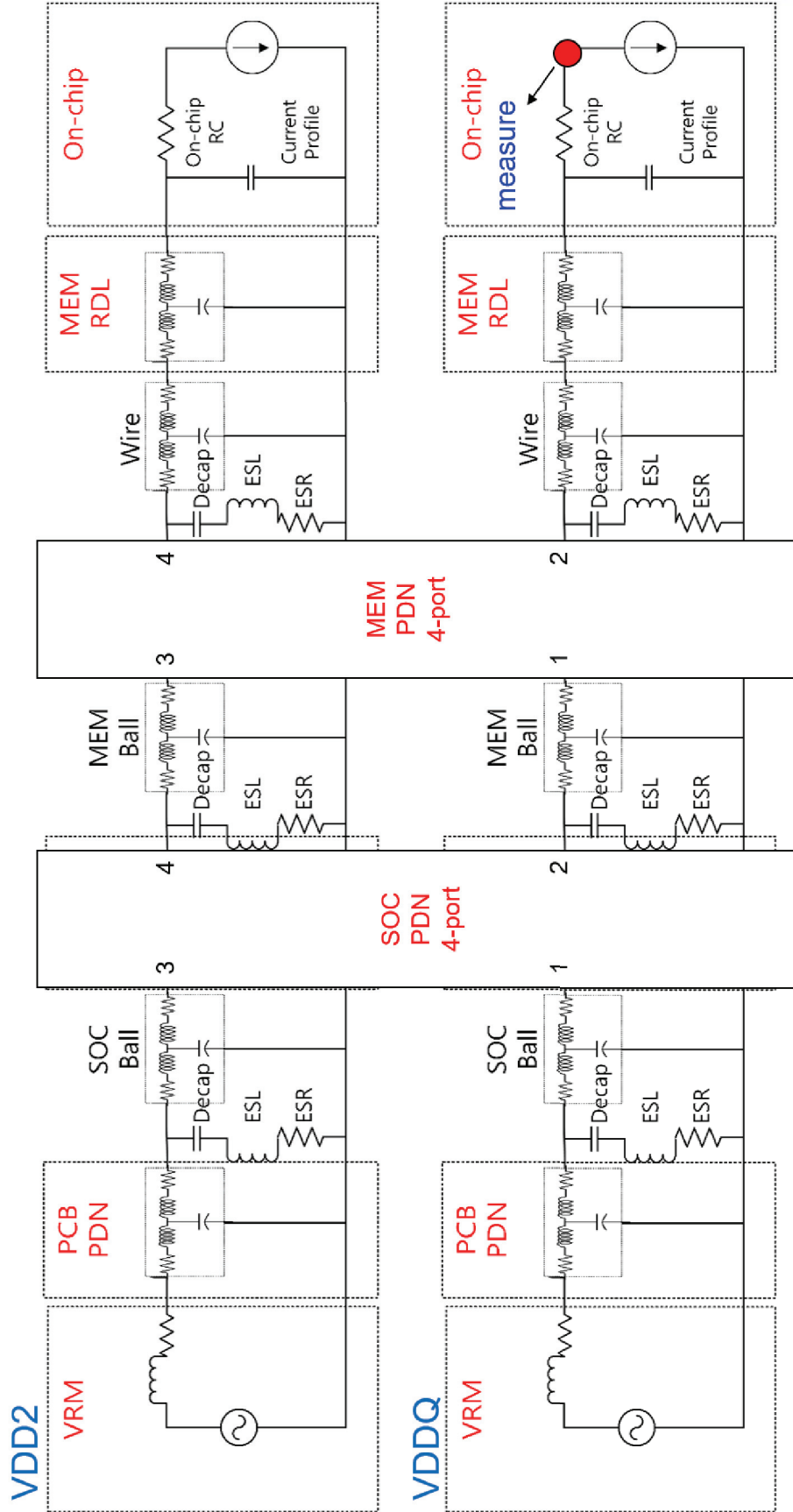


Figure 5.8: Schematic of PoP PDS model and measurement point for on-chip PI analysis.

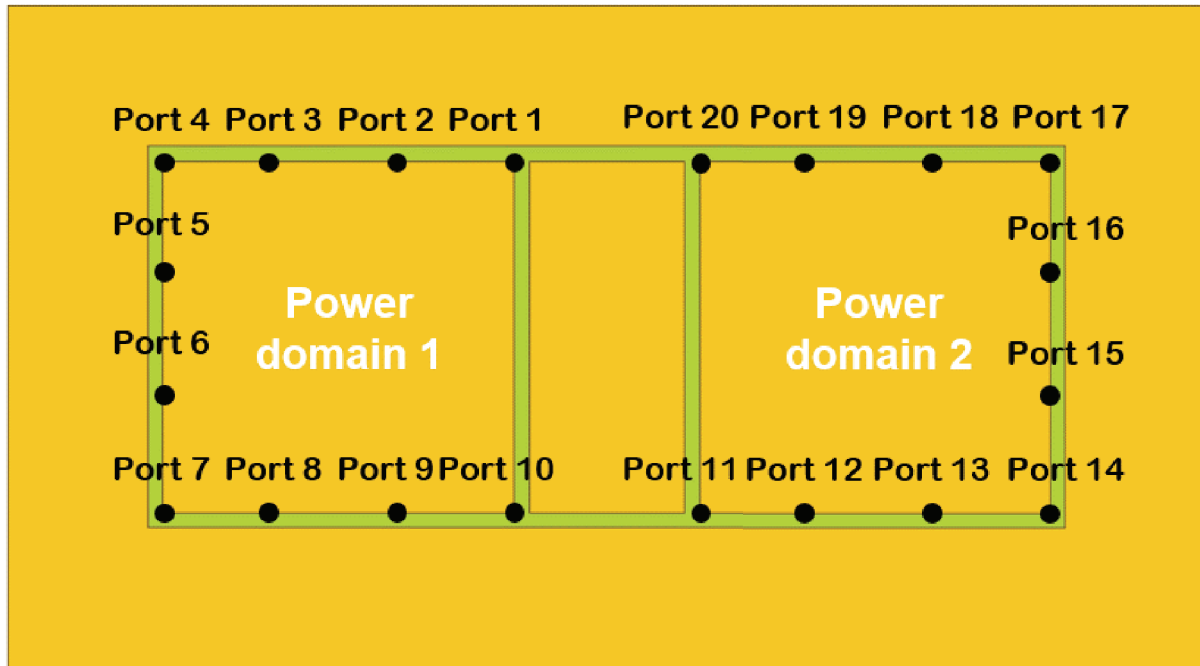


Figure 5.9: Locations of twenty ports for the simulations of PDN.

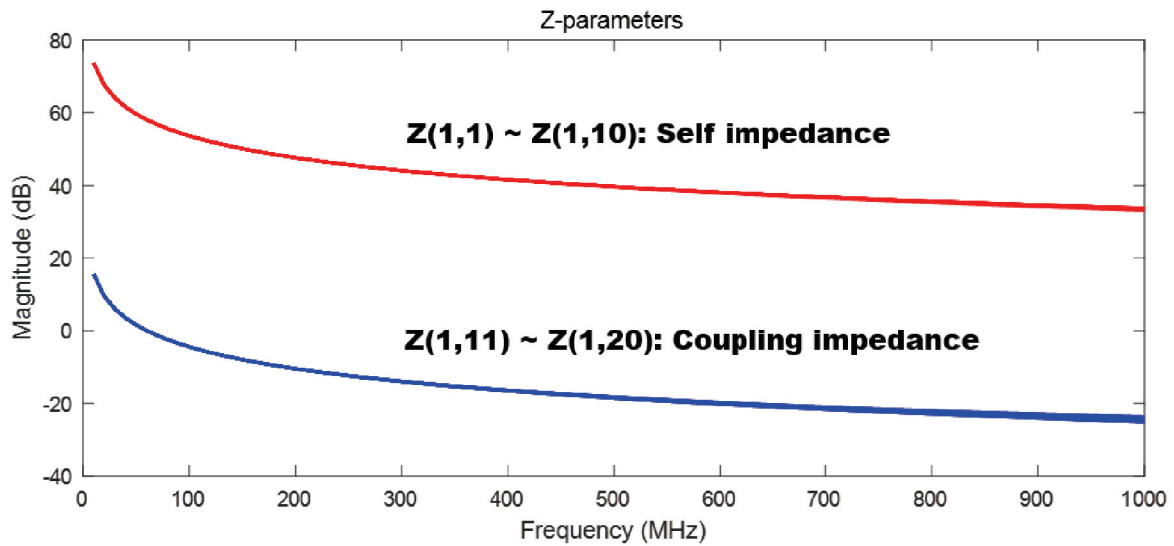


Figure 5.10: Z-parameters of a 20-port PDN structure: CGND = 1, RGND = 1, MARGIN_WIDTH = 2000 μm , CORE_THICKNESS = 400 μm , PSR_THICKNESS = 200 μm .

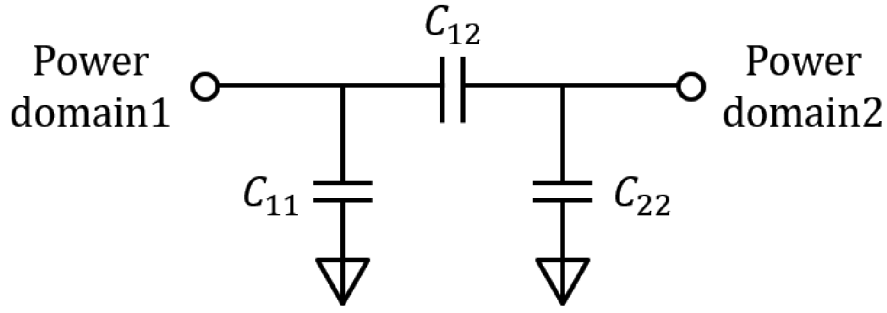


Figure 5.11: 2-port network system of PDN consist of capacitors.

From the Z -parameter data, we can calculate the capacitance $C_{1,1}$ and $C_{1,11}$ from $Z_{1,1}$ and $Z_{1,11}$ by equation 5.8.

$$C_{1,1} = \frac{1}{j\omega Z_{1,1}}, \quad C_{1,11} = \frac{1}{j\omega Z_{1,11}} \quad (5.8)$$

To reduce the effect of numerical errors in the Z -parameter data, we sampled the values of $C_{1,1}$ and $C_{1,11}$ at 10 MHz, 0.5 GHz, and 1 GHz and used the averaged values. By representing the equivalent circuit in Figure 5.11 in terms of Z -parameter elements, the following relations can be found:

$$C_{1,1}^* = C_{11} + C_{12} \parallel C_{22} = \frac{C_{11}C_{12} + C_{11}C_{22} + C_{12}C_{22}}{C_{12} + C_{22}}$$

$$C_{1,11}^* = \frac{C_{11} + C_{22}}{C_{12}} [C_{11} \parallel C_{22} + C_{12}] = \frac{C_{11}C_{12} + C_{11}C_{22} + C_{12}C_{22}}{C_{12}}, \quad (5.9)$$

where $C_{1,1}^*$ is the average value of three $C_{1,1}$ values and $C_{1,11}^*$ is the average value of three $C_{1,11}$. From equation 5.9, we can represent $C_{1,1}^*$ and $C_{1,11}^*$ as C_{11} , C_{12} and C_{22} in the 2-port network system of Figure 5.11. Since power domain 1 and power domain 2 have the identical structure and dimensions, we can assume that C_{11} equals C_{22} . In this condition, we obtain the following simplified relations:

$$C_{11} = \frac{C_{1,1}^* C_{1,11}^*}{C_{1,1}^* + C_{1,11}^*}, \quad C_{12} = \frac{C_{1,1}^{*2} C_{1,11}^*}{C_{1,11}^* - C_{1,1}^{*2}} \quad (5.10)$$

It is important to note that C_{12} is related to the coupling between power domain 1 and power domain 2. A larger C_{12} value results from a smaller $C_{1,11}^*$ value, which comes from a higher magnitude of $Z(1,11)$. Therefore, we can quantify the level of domain coupling by observing the value of C_{12} as well as the magnitude of $Z(1,11)$.

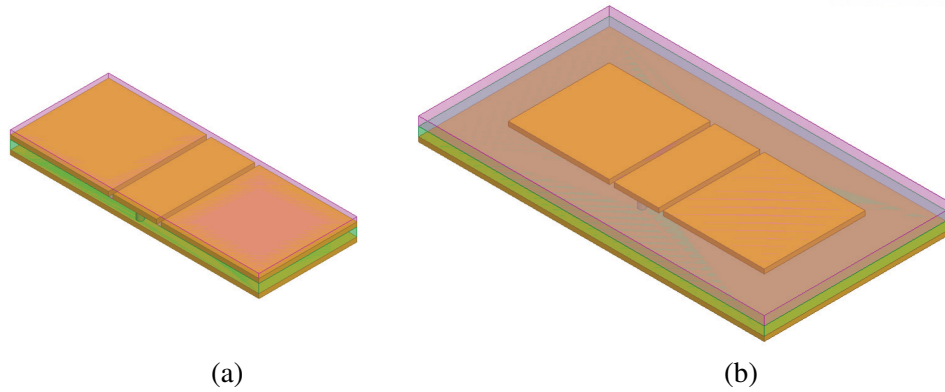


Figure 5.12: PDN structures according to existence and nonexistence of ground margin: (a) structure of PDN that has no ground margin (MARGIN_WIDTH = 0 μm), (b) structure of PDN that has ground margin (MARGIN_WIDTH = 2000 μm).

Effect of ground margin

To verify the effect of ground margin on the power domain coupling, we simulated and analyzed by varying MARGIN_WIDTH from 0 μm to 3,000 μm with the other variables in the modeling parameters are fixed: CGND = 1, RGND = 0, CORE_THICKNESS = 400 μm , PSR_THICKENSS = 200 μm as shown in Figure 5.12. The computed capacitances from the numerical analyses are summarized in Table 5.4, where the power domain coupling gradually decreases as MARGIN_WIDTH increases. This simple parametric analysis shows that the effect of the edge-side margin is significant. The effect can be explained by the role of the bottom ground, which confines the fields from each power domain into the bottom side and reduces the fringing field going to the other power domain.

Effects of upper ground planes

To investigate the effects of the center and the ring ground planes on the power domain coupling, we analyzed three cases: RGND = 1 and CGND = 1, RGND = 0 and CGND = 1, RGND = 1 and CGND = 0 as shown in Figure 5.13. The other variables in the modeling parameters are fixed: MARGIN_WIDTH = 2000 μm , CORE_THICKNESS = 400 μm , PSR_THICKENSS = 200 μm . By comparing the three cases summarized in Table 5.5, we can find that both the ring ground plane and the center ground plane decrease the power domain coupling. Since the main coupling fields exist between the two facing sides of power planes, the center ground plane reduces the coupling more effectively than ring ground plane.

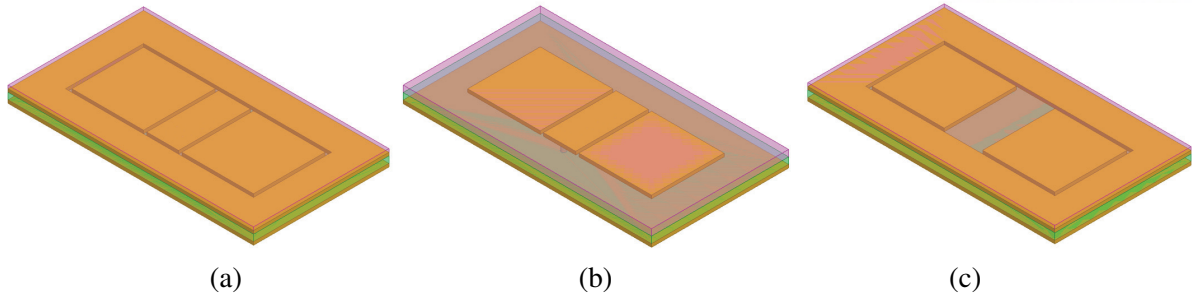


Figure 5.13: PDN structure according to existence and nonexistence of ring ground plane and center ground plane: (a) structure of PDN that has both ring and center ground plane ($RGND = 1$ and $CGND = 1$), (b) structure of PDN that has center ground plane and no ring ground plane ($RGND = 0$ and $CGND = 1$), (c) structure of PDN that has ring ground plane and no center ground plane ($RGND = 1$ and $CGND = 0$).

Effect of substrate thickness

To verify $PSR_THICKNESS$ and $CORE_THICKNESS$ effect about the power domain coupling, we simulated and analyzed three cases: $PSR_THICKNESS = 200\ \mu m$ and $CORE_THICKNESS = 400\ \mu m$, $PSR_THICKNESS = 400\ \mu m$ and $CORE_THICKNESS = 400\ \mu m$, $PSR_THICKNESS = 200\ \mu m$ and $CORE_THICKNESS = 800\ \mu m$. And the other variables in the modeling parameters are fixed: $MARGIN_WIDTH = 2000\ \mu m$, $RGND = 0$, $CGND = 1$. As described in Table 5.6, we can observe that the power domain coupling increases as $PSR_THICKNESS$ increases and also $CORE_THICKNESS$ increases. It is because the increased thickness, especially the core substrate thickness in our case, enables more fringe fields to be formed between domains.

5.3.2 Model verification

We verify our proposed methodology using the PoP PDS model with the *SPICE* simulation results of the full industrial layout¹ under the JEDEC LPDDR4 environment. The main parameters used in the experiment are shown in Table 5.3, and the core (VDDQ) and I/O (VDD2) peak-to-peak ripple voltage are compared in the 800 MHz IC switching-frequency environment. In this experiment, we extract the lumped RLC values of PCB and RDL. In addition, we extract the four-port *S*-parameters of SoC and memory PDN. The lumped RLC values of the ball and wire are calculated based on the real structure, and the same current profile is used for comparison. Table 5.7 shows that the core and I/O peak-to-peak ripple voltages are 1.40% and 0.82%, respectively, for the full layout *SPICE* simulation, 1.61% and 0.73%, respectively for our model with respect to the reference voltage (1.1V). The main cause of the

¹The full layout design is not made public for security reasons; therefore, we received comparison simulation results as the peak-to-peak ripple voltage values for industrial design from the *Samsung Package Development Team*.

error is the difference in considering the parasitic RLC in the model. However, the acceptable error range with the consideration of the reference VDD is 1.1V, and the generally permissible peak-to-peak ripple voltage is 4-10%. Thus, using our proposed methodology, it is possible to analyze the PI using parametric link simulations with a runtime that is 1,000 times faster.

5.3.3 Domain coupling

To investigate the domain coupling effect of the ring and the central ground plane, we simulate two methods. First, we analyze the effect of the existence of the central and ring grounds. Figure 5.13 shows our three cases of the PDN structures in SoC and memory package: RGND = 1 and CGND = 1, RGND = 1 and CGND = 0, and RGND = 0 and CGND = 1. Figure 5.14(a) is the generated current profile. As shown in Figure 5.14(b), both the central and ring grounds can reduce the domain-coupling effect by reducing the mutual capacitance between the two domains. In this case, the central ground plane improves PI more effectively than the ring ground plane. Second, we vary the width of the ground plane margin from 0 μm to 3,000 μm with the other parameters remaining fixed: CGND = 1, RGND = 0. As shown in Figure 5.14(c), we observe that the large margin width can reduce the IR drop. Note that our multi-domain PDN model is relatively smaller than the real-package PDN structure. Therefore, the effect of the domain coupling is also less than the realistic environment.

5.3.4 On-chip decap effect

In general, decap is used to improve PI in various PDS. However, a large decap is expensive and changes the resonance frequency of the system. Therefore, we perform sweep simulation by varying the on-chip decap to determine the proper decap ranging from 0.01 nF to 1.28 nF. As shown in Figure 5.15, the large decap effectively reduces the high-frequency VDD ripple of the on-chip pad. Moreover, we investigate the proper size of the decap from the sweep simulation (see Figure 5.16), which considers the resonance frequency of the PDS with the AC simulation. The 0.06 nF on-chip decap used as an industrial reference is a reasonable value because 0.06 nF of the decap effectively reduces the IR drop and produces approximately 250 MHz of resonance frequency for the PDS. The resonance frequency of 250 MHz is small as compared with the IC switching frequency of the experimental environment. However, PI may be worse if other frequency noises overlap the PDS resonance frequency. Therefore, it is necessary to analyze the input VRM noise of various frequencies to investigate the effect of the resonance frequency.

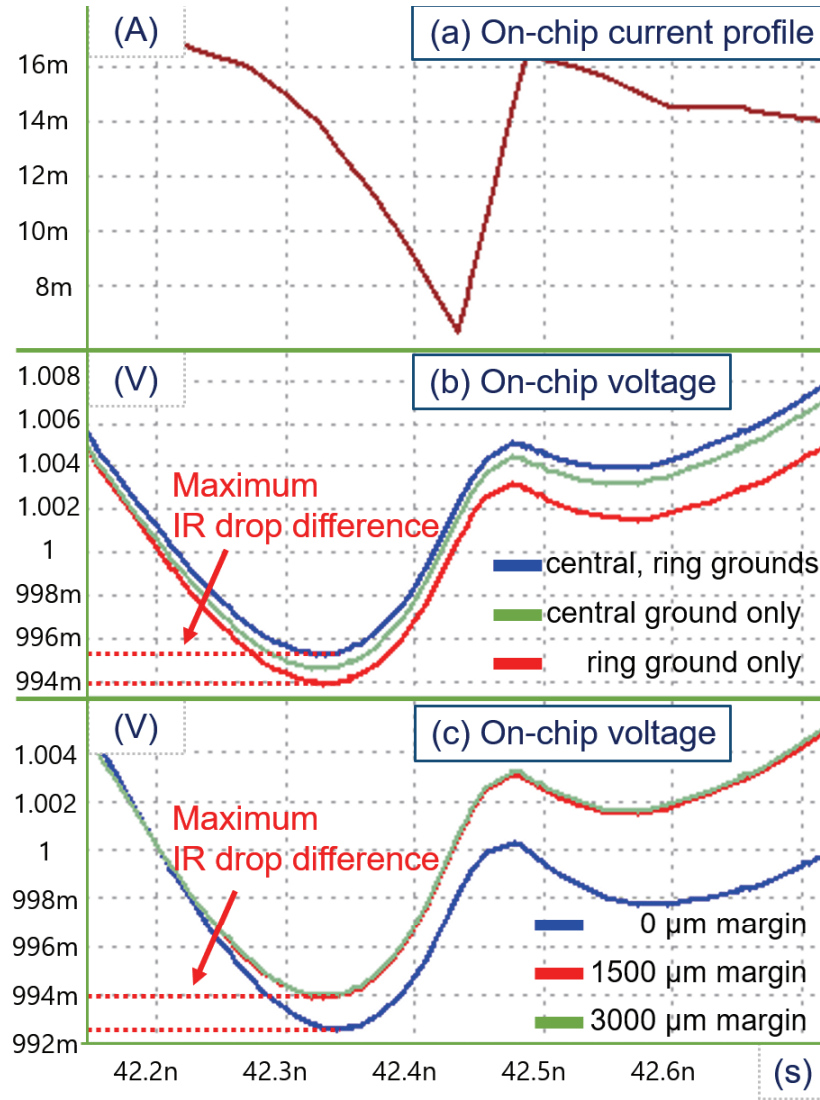


Figure 5.14: Transient results of (a) generated current profile and corresponding voltage fluctuations in two methods (b) voltage with three PDN structures (with both center/ring, without ring, without the central ground) and (c) three ground margins (0 μm , 1500 μm , 3000 μm) of the package PDN on the pad of the on-chip VDDQ.

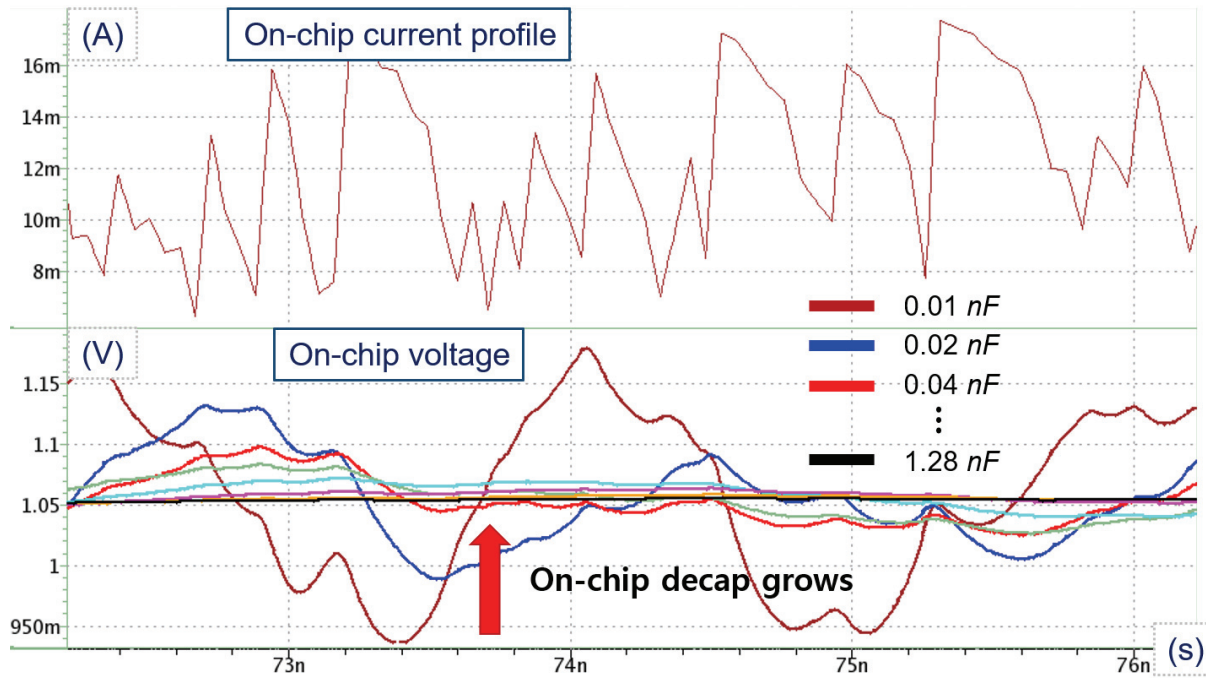


Figure 5.15: Current profile (above) and voltage graph (below) on the pad of the on-chip VDDQ. IR drop on the on-chip pad decreases as the on-chip decap increases.

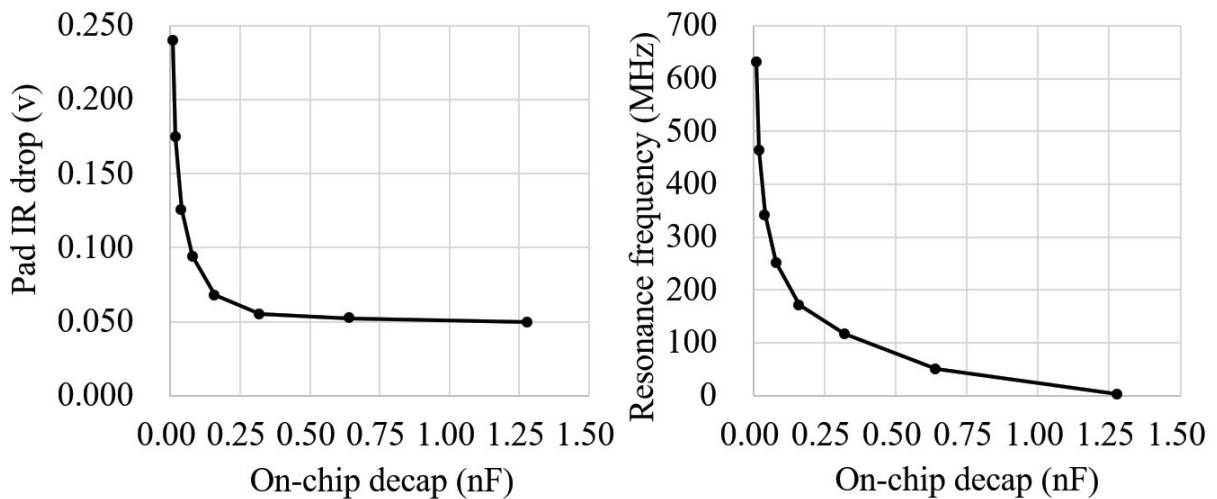


Figure 5.16: IR drop (left) and resonance frequency (right) on the pad of on-chip with various values of the on-chip decap

5.3.5 Input noise effect

To analyze the effect of the input VDD noise frequency, we use a simulation with our proposed methodology. Although the magnitude of the input noise is the same, the PI depends on the frequency. Generally, low/middle-range frequency noise is effectively controlled by on-chip decaps of PCB and packages. However, in this experiment, the decaps are ignored to observe the clear difference of the on-chip VDD fluctuation. In this experiment, we sweep the input noise frequency of VDDQ with the same VDD offset. As shown in Figure 5.17, the effect of the input noise is based on the frequency of the VDDQ of the on-chip pad. In addition, we can analyze the trend of the relationship between the IR drop and the input VDDQ noise frequency, as shown in Figure 5.18(a). Moreover, we analyze how the critical noise frequency range (see the red circle in Figure 5.18(a)) affects the PI. We perform 1,000 Monte Carlo simulations with 10% 1-sigma variation, as shown in Figure 5.18. The total simulation runtime is approximately 120 min, which is significantly faster than the traditional full-layout simulations. As stated in Subsection 5.3.4, the PDS has a resonance frequency of 250 MHz, and the on-chip IR-drop is the worst when the input noise frequency is equal to the resonance frequency (see Figure 5.18). Therefore, the proposed methodology can be a design guideline for a specific input noise, such as determining the appropriate on-chip decap size based on the VRM input noise or changing the RDL structure based on the results of the on-chip simulations.

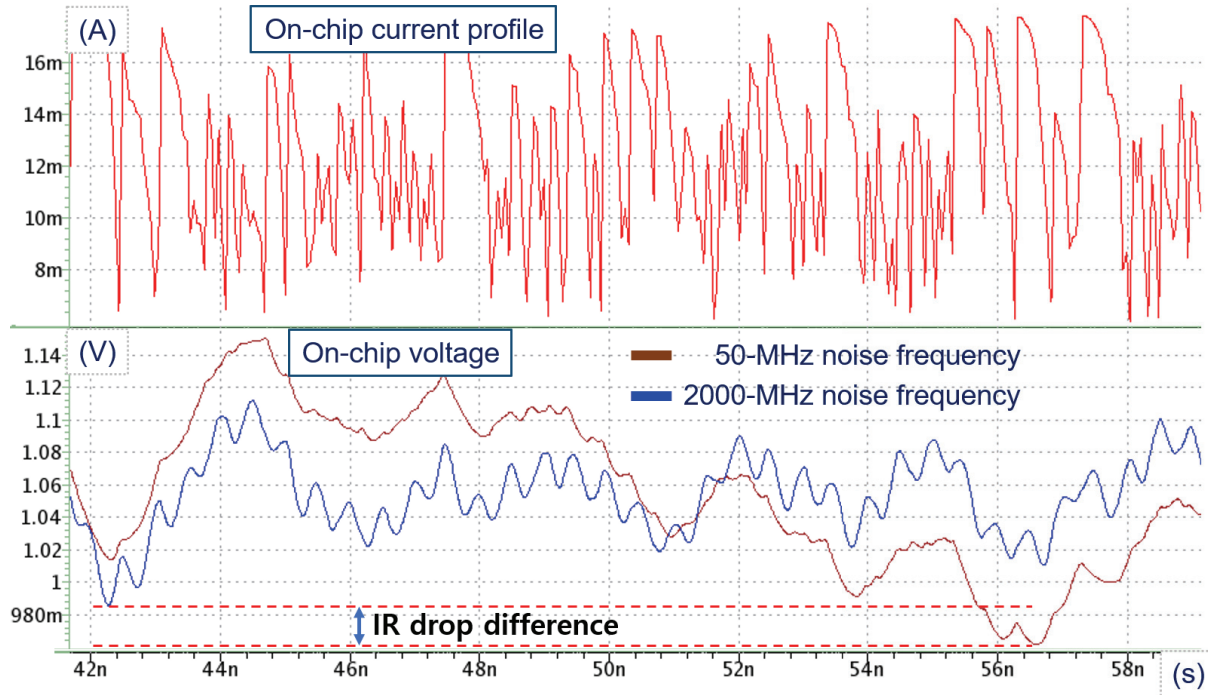


Figure 5.17: Transient result of the voltage (above) and the current profile (below) on the pad of the on-chip VDDQ.

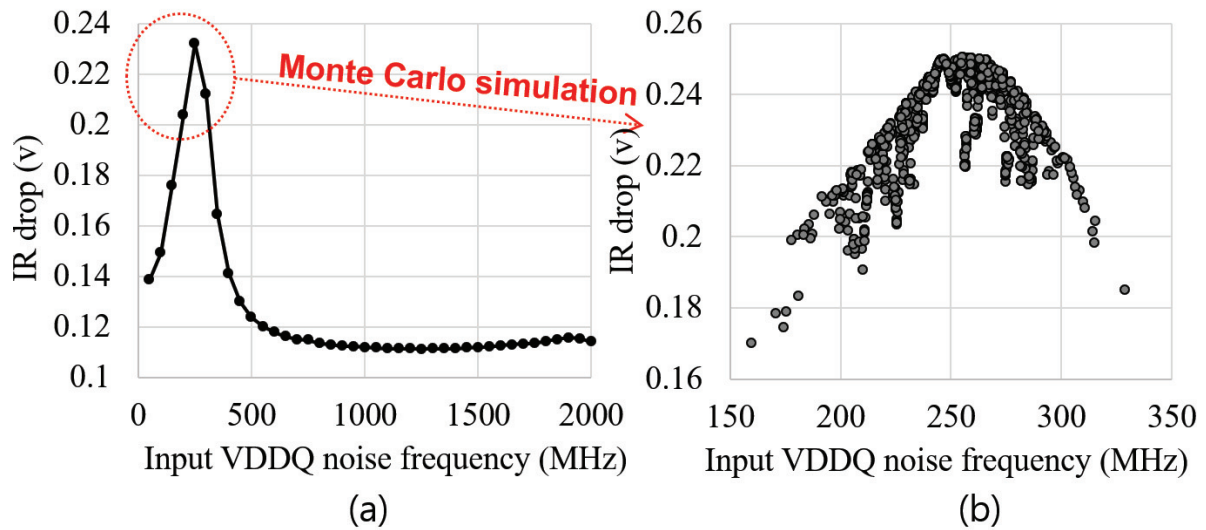


Figure 5.18: IR drop on the on-chip pad with various input noise frequencies; (a) sweep simulations with 100 MHz steps, (b) Monte Carlo simulations.

5.4 Conclusions and Future Directions

In this chapter, we propose a fast PI coanalysis methodology for a multi-domain high-speed memory system. Our methodology has several advantages over the conventional full-layout *SPICE*-based analysis and the case-specific design model-based analysis. First, our methodology improves the simulation flexibility and the runtime with a pseudo-randomized current profile generation that characterizes the realistic switching current. Second, our PI analysis is fast and has massive parametric link simulations. Third, we consider the structural changes of PDN using multiple power domains to analyze the domain coupling effect. We compare our PDS model and our proposed methodology with the results of the industrial full-layout *SPICE* simulation. Our experimental studies confirm the PI improvement by the structural changes of the ground plane in a multiple power domain PDN. Moreover, we determine an proper on-chip decap value based on the resonance frequency under the LPDDR4 800 MHz environment by using fast Monte Carlo analysis. Thus, our proposed PI coanalysis methodology can be used by high-speed memory package designers as a design guideline to predict PI.

5.5 Acknowledgments

Chapter V is extension of “Fast Chip-Package-PCB Coanalysis Methodology for Power Integrity of Multi-domain High-Speed Memory: A Case Study”, *Proc. IEEE/ACM Design, Automation and Test, in Europe* 2018; and “Power Integrity Coanalysis Methodology for Multi-Domain High-Speed Memory Systems” *IEEE Access* (2019) to appear. The multiple power domain PDN model used in this chapter refers to our preliminary study, “A Preliminary Analysis of Domain Coupling in Package Power Distribution Network”, *Proc. IEEE International Symposium on Radio-Frequency Integration Technology*, 2017.

I would like to thank my coauthors Professor Ki Jin Han, Professor Youngmin Kim, and Professor Seokhyeong Kang.

Table 5.1: Modeling parameters and dimensions of multi-domain PDN. The value is the parameter value that is the reference of the PDN structure to be used in the simulation.

Name	Description	Value
PAD_PITCH	Interval length of ports	$1500\mu m$
PLANE_SPACING	Interval length of planes in a same layer	$200\mu m$
PLANE_THICKNESS	Thickness of the power domain planes and ground planes	$200\mu m$
POWER_DOMAIN 1_X,Y	The x-y-dimensions of the power domain 1 plane	$4500\mu m$
POWER_DOMAIN 2_X,Y	The x-y-dimension of the power domain 2 plane	$4500\mu m$
CGND	Central ground plane	1 (exists) or 0 (not)
CGND_X	The x dimension of the central ground plane	$2000\mu m$
CGND_Y	The y dimension of the central ground plane	$4500\mu m$
RGND	Ring ground plane	1 (exists) or 0 (not)
MARGIN_WIDTH	Width of the extra part outside power domain planes in the edge of PDN	Varying
CORE	Dielectric	$\epsilon_r = 3.9$, $\tan\delta = 0.02$
CORE_THICKNESS	Thickness of dielectric	$400\mu m$
PSR	Prepreg	$\epsilon_r = 3.9$, $\tan\delta = 0.029$
PSR_THICKNESS	Thickness of prepreg	$200\mu m$
VIA	Ground vias	$d = 300\mu m$
VIA_PITCH	Interval length of ground vias	$500\mu m$

Table 5.2: Sinusoidal source parameters.

Sinusoidal parameter	Units
V_0 - offset voltage	volt
V_a - peak amplitude of voltage	volt
$Freq$ - frequency	hertz
T_d - delay time	second
D_f - damping factor	1/second

Table 5.3: Values of the used parameters that significantly affect the simulation results.

Parameter types	Names	Nominal values	Units
General model	On_Chip_R	20	mohm
	On_Chip_C	0.06	nF
Current profile	Time_length	80	ns
	Delay_VDDQ	20	ns
	Interval_VDDQ	0	ps
	Slope_step_MIN/MAX_VDDQ	50/60	ps
	I_MIN/MAX_VDDQ	6/18	mA
	Delay_VDD2	20	ns
	Interval_VDD2	0	ps
	Slope_step_MIN/MAX_VDD2	50/100	ps
	I_MIN/MAX_VDD2	9/50	mA
	VDDQ_nom	1.1	V
VRM	VDDQ_noise	0	V
	VDDQ_freq	500	MHz
	VDD2_nom	1.1	V
	VDD2_noise	0.011	V
	VDD2_freq	500	MHz

Table 5.4: Margin effect about the power domain coupling.

MARGIN_WIDTH	$C_{1,1}^*$ (pF)	$C_{1,11}^*$ (pF)	C_{11} (pF)	C_{12} (fF)
0 μm	2.369	157.0	2.334	35.75
500 μm	2.730	337.1	2.708	22.11
1000 μm	2.798	577.4	2.785	13.56
1500 μm	2.807	868.8	2.798	9.07
2000 μm	2.819	1090.9	2.812	7.28
2500 μm	2.818	1215.0	2.811	6.54
3000 μm	2.820	1377.5	2.814	5.77

Table 5.5: Ground plane effect about the power domain coupling

Ground plane	$C_{1,1}^*$ (pF)	$C_{1,11}^*$ (pF)	C_{11} (pF)	C_{12} (fF)
RGND = 1, CGND = 1	3.383	2746	3.379	4.168
RGND = 0, CGND = 1	2.819	1091	2.812	7.280
RGND = 1, CGND = 0	3.194	1060	3.184	9.624

Table 5.6: PSR_THICKNESS and CORE_THICKNESS effect about the power domain coupling

Ground plane	$C_{1,1}^*$ (pF)	$C_{1,11}^*$ (pF)	C_{11} (pF)	C_{12} (fF)
CORE = 400 μm , PSR = 200 μm	2.819	1091	2.812	7.28
CORE = 400 μm , PSR = 400 μm	2.887	981.1	2.879	8.495
CORE = 800 μm , PSR = 200 μm	1.848	326.5	1.838	10.460

Table 5.7: Comparison of peak-to-peak ripple voltage results obtained by full layout *SPICE* simulation and proposed methodology.

Peak-to-peak ripple voltage		
	Full layout <i>SPICE</i> (mV)	Proposed (mV)
Core	15.41	17.72
I/O	9.0	8.0
	Full layout <i>SPICE</i>	Proposed
Average Runtime	7200s	7.0s

Chapter VI

Conclusion and Future Consideration

This chapter summarizes key contributions of this dissertation and presents future directions for optimization of power delivery in future VLSI designs.

A combined three-dimensional (3D) integration technique and a heterogeneous architectural structure is emerging as a promising solution to overcome Moore's law in the modern VLSI designs. The 3D heterogeneous architectural structure is growing attention because it reduces costs and time-to-market by increasing manufacturing yield with high integration rate and modularization. However, a main design concern of heterogeneous 3D architectural structure is power management for lowering power consumption with maintaining the required power integrity from IR drop. Although the low-power design can be realized in front-end-of-line level by reduced power supply complementary metal–oxide–semiconductor technologies, the overall low-power system performance is available with a proper design of power delivery network (PDN) for chip-level modules and system-level architectural structure. Thus, there is a demand for both the coanalysis and optimization for both chip-level and system-level. We have been analyzed and optimized power delivery on-chip in various 3D integration environments, and we also have proposed a chip-package-PCB coanalysis methodology at the system level.

We first have proposed a novel power gating control scheme in the through-silicon-via (TSV)-based 3D IC. We investigate the in-rush current in the power gating of a 3D IC coupled with frequency-dependent tapered TSV models and propose adaptive power gating technique to maximize performance by reducing wake-up time. This study enables to prevent performance loss due to pessimistic prediction in two or more multi-layer die stacking environment.

Second, as the size and the length of the inter-tier vias (VIs) become very smaller, the density of the circuit increases but the IR drop issue has emerged due to higher resistance of vertical interconnection more than TSVs. Moreover, power delivery network (PDN) competes with signal routing VIs for

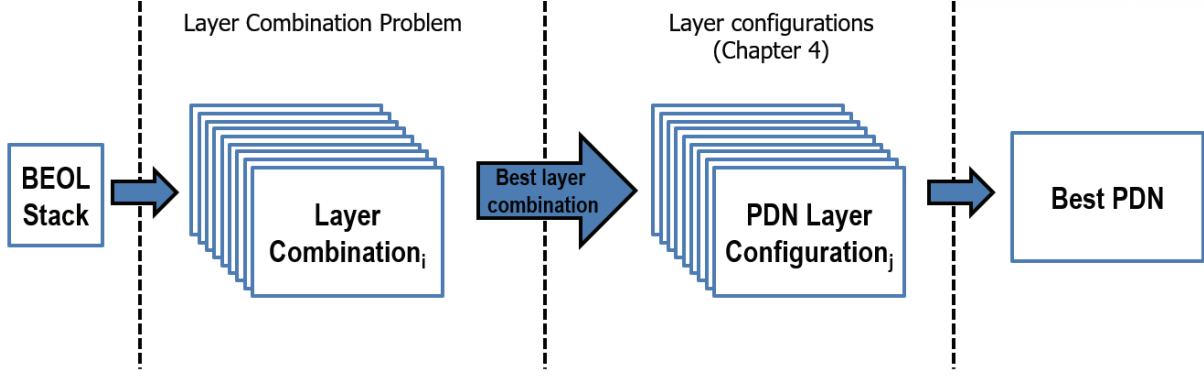


Figure 6.1: High-level overall flow for identification of best PDN.

restricted budget of resources and cost, these competitions may cause breakdown of design implementation and performance degradation. Thus, we propose a machine learning based PDN optimization methodology for emerging D2W integration to identify an nearly optimal PDN for a given design and PDN specification. To the best of our knowledge, we are the first to propose such a pathfinding methodology to identify nearly-optimal PDN configurations for D2W-based designs.

Third, we extend the observation to system-level, we have proposed a power integrity coanalysis methodology for multiple power domains in high-frequency memory systems. Our coanalysis methodology can analyze the tendencies in power integrity by using parametric methods with consideration of package-on-package integration. We have proved that our methodology can predict similar peak-to-peak ripple voltages that are comparable with the realistic simulations of high-speed low-power memory interfaces.

Finally, we propose analysis and optimization methodologies that are generally applicable to various integration methods used in modern VLSI designs as computer-aided-design-based solutions.

Our current special research interest is an extension of PDN pathfinding. The design-specific PDN choices at the “*Pareto frontier*” of IR drop versus routability are not addressed even in 2D ICs yet. Because the number of available metal layers of BEOL is high in advanced technology nodes, we need to choose which layer combination among multiple possible layers to construct. However, PDN design solution space for a given BEOL is large (e.g., layer combination \times layer configuration). Thus, decoupling layer combination from layer configuration can help us efficiently explore the overall PDN solution space as shown in Figure 6.1. Once the best layer combination is found, we can then apply our optimization methodology (in Chapter IV) to find the best layer configuration. Finally, we seek to find optimal PDN among possible PDN combination as shown in Figure 6.2. In addition, we extend our approach to integration technologies other than face-to-face integration.

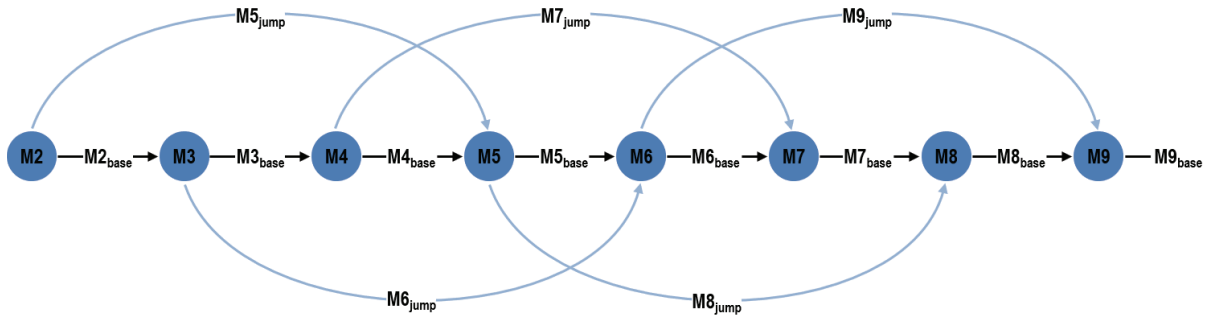


Figure 6.2: An example of a graph of the possible PDN combination cases for a total of nine BEOL layers. Metal layer 2 is assumed to be a power rail.

Bibliography

- [1] A. V. Mezhiba and E. G. Friedman, “Scaling Trends of On-Chip Power Distribution Noise”, *IEEE Transactions on Very Large Scale Integration* 12(4) (2004), pp. 386-394.
- [2] G. Yeap, “Smart Mobile SoCs Driving the Semiconductor Industry: Technology Trend, Challenges and Opportunities”, *Proc. IEDM*, 2013, pp. 1-3.
- [3] Q. Ma and H. Fujimoto, “Silicon Interposer and Multi-Chip-Module (MCM) with Through Substrate Vias”, *U.S. Patent* No.6,229,216, 2001.
- [4] K. Lee and A. Barber, “Modeling and Analysis of Multichip Module Power Supply Planes”, *IEEE Transactions on Components, Packaging & Manufacturing Technology* 18(4), (1995), pp. 628-639.
- [5] P. E. Garrow and I. Turlik, *Multichip Module Technology Handbook* (Vol. 688) New York: McGraw-Hill, 1998.
- [6] F. Norrod, “Evolving System Design for HPC: Next Generation CPU and Accelerator Technologies”, *Proc. OGHPC*, 2019.
- [7] G. Mounce, J. Lyke, S. Horan, W. Powell, R. Doyle and R. Some, “Chiplet based Approach for Heterogeneous Processing and Packaging Architectures”, *Proc. AeroConf*, 2016, pp. 1-12.
- [8] J. H. Lau, *Heterogeneous Integrations*, Springer, 2019.
- [9] M. Swaminathan and E. Engin, *Power Integrity Modeling and Design for Semiconductors and Systems*, Prentice Hall, 2007.
- [10] G. Huang, M. S. Bakir, A. Naeemi and J. D. Meindl, “Power Delivery for 3D Chip Stacks: Physical Modeling and Design Implication”, *IEEE Transactions on Components, Packaging & Manufacturing Technology* 2(5) (2012), pp. 852-859.

- [11] M. B. Healy and S. K. Lim, "Power Delivery System Architecture for Many-Tier 3D Systems", *Proc. ECTC*, 2010, pp. 1682-1688.
- [12] M. Jung and S. K. Lim, "A Study of IR-Drop Noise Issues in 3D ICs with Through-Silicon-Vias", *Proc. 3DIC*, 2010, pp. 1-7.
- [13] S. Q. Gu, P. Marchal, M. Facchini, F. Wang, M. Suh, D. Lisk and M. Nowak, "Stackable Memory of 3D Chip Integration for Mobile Applications", *Proc. IEDM*, 2008, pp. 1-4.
- [14] D. H. Kim, S. Mukhopadhyay and S. K. Lim, "TSV-aware Interconnect Length and Power Prediction for 3D Stacked ICs", *Proc. IITC*, 2009, pp. 26-28.
- [15] G. Van der Plas, P. Limaye, I. Loi, A. Mercha, H. Oprins, C. Torregiani, S. Thijs, D. Linten, M. Stucchi, G. Katti, D. Velenis, V. Cherman, B. Vandeveld, V. Simons, I. De Wolf, R. Labie, D. Perry, S. Bronckers, N. Minas, M. Cupac, W. Ruythooren, J. Van Olmen, A. Phommahaxay, M. de Potter de ten Broeck, A. Opdebeeck, M. Rakowski, B. De Wachter, M. Dehan, M. Nelis and R. Agarwal, "Design Issues and Considerations for Low-Cost 3-D TSV IC Technology", *IEEE Journal of Solid-State Circuits* 46(1) (2011), pp. 203-307.
- [16] S. Shigematsu, S. I. Mutoh, Y. Matsuya, Y. Tanabe and J. Yamada, "A 1-V High-Speed MTCMOS Circuit Scheme for Power-Down Application Circuits", *IEEE Journal of Solid-State Circuits* 32(6) (1997), pp. 861-869.
- [17] Y. Shin, J. Seomun, K. M. Choi and T. Sakurai, "Power Gating: Circuits, Design Methodologies, and Best Practice for Standard-Cell VLSI Designs", *ACM Transactions on Design Automation of Electronic Systems* 15(4) (2010), pp. 28-37.
- [18] Z. Hu, A. Buyuktosunoglu, V. Srinivasan, V. Zyuban, H. Jacobson and P. Bose, "Microarchitectural Techniques for Power Gating of Execution Units", *Proc. ISLPED*, 2004, pp. 32-37.
- [19] K. Agarwal, K. Nowka, H. Deogun and D. Sylvester, "Power Gating with Multiple Sleep Modes", *Proc. ISQED*, 2006, pp. 633-637.
- [20] H. Singh, K. Agarwal, D. Sylvester and K. J. Nowka, "Enhanced Leakage Reduction Techniques using Intermediate Strength Power Gating", *IEEE Transactions on Very Large Scale Integration* 15(11) (2007), pp. 1215-1224.
- [21] M. H. Chowdhury, J. Gjanci and P. Khaled, "Innovative Power Gating for Leakage Reduction", *Proc. ISCAS*, 2008, pp. 1568-1571.

- [22] Z. Zhang, X. Kavousianos, K. Chakrabarty and Y. Tsiatouhas, “A Robust and Reconfigurable Multi-Mode Power Gating Architecture”, *Proc. VLSI Design*, 2011, pp. 280-285.
- [23] A. B. Kahng, S. Kang, T. S. Rosing and R. Strong, “Many-Core Token-Based Adaptive Power Gating”, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 32(8) (2013), pp. 1288-1292.
- [24] Y. Wang, J. Xu, Y. Xu, W. Liu and H. Yang, “Power Gating aware Task Scheduling in MPSoC”, *IEEE Transactions on Very Large Scale Integration* 19(10) (2011), pp. 1801-1812.
- [25] M. C. Lee, Y. Shi and S. C. Chang, “Efficient Wakeup Scheduling Considering Both Resource Usage and Timing Budget for Power Gating Designs”, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 31(7) (2012), pp. 1041-1049.
- [26] C. Xu, H. Li, R. Suaya and K. Banerjee, “Compact AC Modeling and Performance Analysis of Through-Silicon Vias in 3-D ICs”, *IEEE Transactions on Electronic Devices* 57(12) (2010), pp. 3405-3417.
- [27] J. Kim, J. S. Pak, J. Cho, E. Song, J. Cho, H. Kim, T. Song, J. Lee, H. Lee, K. Park, S. Yang, M.-S. Suh, K.-Y. Byun and J. Kim, “High-Frequency Scalable Electrical Model and Analysis of a Through Silicon Via (TSV)”, *IEEE Transactions on Components, Packaging & Manufacturing Technology* 1(2) (2011), pp. 181-195.
- [28] N. H. Khan, S. M. Alam and S. Hassoun, “System-Level Comparison of Power Delivery Design for 2D and 3D ICs”, *Proc. 3DIC*, 2009, pp. 1-7.
- [29] H. He and J. Q. Lu, “Compact Models of Voltage Drops in Power Delivery Network for TSV-Based Three-Dimensional Integration”, *IEEE Electron Device Letters* 34(3) (2013), pp. 438-440.
- [30] A. E. Ruehli, “Equivalent Circuit Models for Three-Dimensional Multiconductor Systems”, *IEEE Transactions on Microwave Theory and Techniques* 22(3) (1974), pp. 216-221.
- [31] S. Kim, K. J. Han, S. Kang and Y. Kim, “Analysis and Reduction of Voltage Noise of Multi-Layer 3D IC with PEEC-Based PDN and Frequency-Dependent TSV Models”, *Proc. ISOCC*, 2014, pp. 124-125.
- [32] K. J. Han, M. Swaminathan and T. Bandyopadhyay, “Electromagnetic Modeling of Through-Silicon Via (TSV) Interconnections using Cylindrical Modal basis Functions”, *IEEE Transactions on Antennas and Propagation* 33(4) (2010), pp. 804-817.

- [33] K. J. Han and M. Swaminathan, "Inductance and Resistance Calculations in Three-Dimensional Packaging using Cylindrical Conduction-Mode basis Functions", *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 28(6) (2009), pp. 846-859.
- [34] W. H. Lee, S. Pant and D. Blaauw, "Analysis and Reduction of On-Chip Inductance Effects in Power Supply Grids", *Proc. ISSCS*, 2004, pp. 131-136.
- [35] S. Kim, S. Kang, K. J. Han and Y. Kim, "Novel Adaptive Power Gating Strategy of TSV-Based Multi-Layer 3D IC", *Proc. ISQED*, 2015, pp. 537-541.
- [36] A. Todri, S. Kundu, P. Girard, A. Bosio, L. Dilillo and A. Virazel, "A Study of Tapered 3-D TSVs for Power and Thermal Integrity", *IEEE Transactions on Very Large Scale Integration* 21(2) (2012), pp. 306-319.
- [37] K. Athikulwongse, A. Chakraborty, J.-S Yang, D. Z. Pan, S. K. Lim, "Stress-Driven 3D-IC Placement with TSV Keep-Out Zone and Regularity Study", *IEEE Press* (2010), pp. 669-674.
- [38] K. Arabi, K. Samadi and Y. Du, "3D VLSI: A Scalable Integration Beyond 2D", *Proc. ISPD*, 2015, pp. 1-7.
- [39] W. J. Chan, Y. Du, A. B. Kahng, S. Nath, K. Samadi, "3DIC Benefit Estimation and Implementation Guidance from 2DIC Implementation", *Proc. DAC*, 2015, pp. 1-6.
- [40] K. Chang, S. Sinha, B. Cline, R. Southerland, M. Doherty, G. Yeric and S. K. Lim, "Cascade2D: A Design-Aware Partitioning Approach to Monolithic 3D IC with 2D Commercial Tools", *Proc. ICCAD*, 2016, pp. 1-8.
- [41] K. Chang, S. Das, S. Sinha, B. Cline, G. Yeric and S. K. Lim, "Frequency and Time Domain Analysis of Power Delivery Network for Monolithic 3D ICs", *Proc. ISLPED*, 2017, pp. 1-6.
- [42] K. Chang, A. Koneru, K. Chakrabarty and S. K. Lim, "Design Automation and Testing of Monolithic 3D ICs: Opportunities, Challenges, and Solutions", *Proc. ICCAD*, 2017, pp. 805-810.
- [43] Y. Du, K. Samadi and K. Arabi, "Emerging 3DVLSI: Opportunities and Challenges", *Proc. S3S*, 2015, pp. 1-5.
- [44] J. H. Friedman, "Multivariate Adaptive Regression Splines", *The Annals of Statistics* 19(1) (1991), pp. 1-67.

- [45] A. Kahng, A. B. Kahng, H. Lee and J. Li, "PROBE: A Placement, Routing, Back-End-of-Line Measurement Utility", *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 37(7) (2018), pp. 1459-1472.
- [46] M.-C. Kim, J. Hu, D.-J. Lee and I. L. Markov, "A SimPLR Method for Routability-Driven Placement", *Proc. ICCAD*, 2011, pp. 67-73.
- [47] B. W. Ku, P. Debacker, D. Milojevic, P. Raghaven, D. Verkest, A. Thean and S. K. Lim, "Physical Design Solutions to Tackle FEOL/BEOL Degradation in Gate-Level Monolithic 3D ICs", *Proc. ISLPED*, 2016, pp. 76-81.
- [48] B. W. Ku, K. Chang and S. K. Lim, "Compact-2D: A Physical Design Methodology to Build Commercial-Quality Face-to-Face-Bonded 3D ICs", *Proc. ISPD*, 2018, pp. 90-97.
- [49] S. Panth, K. Samadi, Y. Du and S. K. Lim, "Design and CAD Methodologies for Low Power Gate-Level Monolithic 3D ICs", *Proc. ISLPED*, 2014, pp. 171-176.
- [50] S. Panth, K. Samadi, Y. Du and S. K. Lim, "Tier-Partitioning for Power Delivery vs Cooling Tradeoff in 3D VLSI for Mobile Applications", *Proc. DAC*, 2015, pp. 1-6.
- [51] S. Panth, K. Samadi, Y. Du and S. K. Lim, "Shrunk-2D: A Physical Design Methodology to Build Commercial-Quality Monolithic 3D ICs", *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 36(10) (2017), pp. 1716-1724.
- [52] K. Pearson, "Note on Regression and Inheritance in the Case of Two Parents", *Proc. RSL*, 1895, pp. 240-242.
- [53] Y. Peng, D. Petranovic, K. Samadi, P. Kamal, Y. Du and S. K. Lim, "Inter-die Coupling Extraction and Physical Design Optimization for Face-to-Face 3D ICs", *IEEE Transactions on Nanotechnology* (2017), pp. 1-1.
- [54] S. K. Samal, K. Samadi, P. Kamal, Y. Du and S. K. Lim, "Full Chip Impact Study of Power Delivery Network Designs in Gate-Level Monolithic 3-D ICs", *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 36(6) (2017), pp. 992-1003.
- [55] C. Spearman, "The Proof and Measurement of Association between Two Things", *The American Journal Philology on JSTOR* 15(1) (1904), pp. 72-101.
- [56] H. Reiter, "TSMC Details Family of Chip Stacks", https://www.eetimes.com/author.asp?section_id=36&doc_id=1322075

- [57] J. Rudy, py-earth, <https://github.com/scikit-learn-contrib/py-earth>
- [58] G. Yeric, “Three Dimensions in 3DIC - Part I”, <https://community.arm.com/arm-research/b/articles/posts/three-dimensions-in-3dic-part-1>
- [59] G. Yeric, “Three Dimensions in 3DIC - Part II”, <https://community.arm.com/arm-research/b/articles/posts/three-dimensions-in-3dic-part-ii>
- [60] M. LePeduc, “The Chiplet Race Begins”, <https://semiengineering.com/the-chiplet-race-begins/>
- [61] Cadence, “Chiplets – Reinventing Systems Design”, https://community.cadence.com/cadence_blogs_8/b/spi/posts/chiplets
- [62] OpenCores: Open Source IP-Cores, <http://www.opencores.org>
- [63] Cadence Innovus User Guide, <https://www.cadence.com>
- [64] ANSYS RedHawk User Guide, <https://www.ansys.com>
- [65] Synopsys Design Compiler User Guide, <http://www.synopsys.com>
- [66] S. Kim, K. J. Han, Y. Kim and S. Kang, “Fast Chip-Package-PCB Coanalysis Methodology for Power Integrity of Multi-Domain High-Speed Memory: A Case Study”, *Proc. DATE*, 2018, pp. 885-888.
- [67] S. Lee, H. Cho, Y. H. Son, Y. Ro, N. S. Kim and J. H. Ahn, “Leveraging Power-Performance Relationship of Energy-Efficient Modern DRAM Devices”, *IEEE Access* 6 (2018), pp. 31387-31398.
- [68] M. Swaminathan and E. Engin, *Power Integrity Modeling and Design for Semiconductors and Systems*, Prentice Hall, 2007.
- [69] S. L. Huh and H. Shi, “Detection of Noise Coupling between Power Domains on Package”, *Proc. ECTC*, 2016, pp. 2028-2033.
- [70] JEDEC LPDDR4 Standard, <https://www.jedec.org/sites/default/files/docs/JESD209-4.pdf>
- [71] Z. Zeng, X. Ye, Z. Feng and P. Li, “Tradeoff Analysis and Optimization of Power Delivery Networks with On-Chip Voltage Regulation”, *Proc. DAC*, 2010, pp. 831-836.

- [72] D. H. Lee, Y. S. Shin, C. G. Kim, J. H. Song, J. K. Wee, J. M. Lee and J. S. Seol, “Design of Multiple Power Domains based on Ground Separation Technique for Low-Noise and Small-Size Module”, *Proc. APEMC*, 2012, pp. 805-808.
- [73] M. E. Kowalski and P. Codd, “Co-simulation of IC, Package and PCB Power Delivery Networks in Ultra-Low Voltage Power Rail Designs”, *Proc. ECTC*, 2007, pp. 798-803.
- [74] Y. M. Lee and C. C. P. Chen, “The Power Grid Transient Simulation in Linear Time based on 3-D Alternating-Direction-Implicit Method”, *IEEE Transaction on Computer-Aided Design of Integrated Circuits and Systems* 22(11) (2003), pp. 1545-1550.
- [75] M. Hsieh, “Advanced Flip Chip Package on Package Technology for Mobile Applications”, *Proc. ICEPT*, 2016, pp. 486-491.
- [76] D. M. Mathew, M. Schultheis, C. C. Rheinländer, C. Sudarshan, C. Weis, N. Wehn and M. Jung, “An Analysis on Retention Error Behavior and Power Consumption of Recent DDR4 DRAMs”, *Proc. DATE*, 2018, pp. 293-296.
- [77] J. Kim, W. Lee, Y. Shim, J. Shim, K. Kim, J. S. Pak and J. Kim, “Chip-Package Hierarchical Power Distribution Network Modeling and Analysis Based on a Segmentation Method”, *IEEE Transactions on Antennas and Propagation* 33(3) (2010), pp. 647-659.
- [78] J. Feng, B. Dhavale, J. Chandrasekhar, Y. Tretiakov and D. Oh, “System Level Signal and Power Integrity Analysis for 3200Mbps DDR4 Interface”, *Proc. ECTC*, 2013, pp. 1081-1086.
- [79] K. Jeong, A. B. Kahng, S. Kang, T. S. Rosing and R. Strong, “MAPG: Memory Access Power Gating”, *Proc. DATE*, 2012, pp. 1054-1059.
- [80] H.-H. Chuang, W.-D. Guo, Y.-H. Lin, H.-S. Chen, Y.-C. Lu, Y.-S. Cheng, M.-Z. Hong, C.-H. Yu, W.-C. Cheng, Y.-P. Chou, C.-J. Chang, J. Ku, T.-L. Wu and R.-B. Wu, “Signal/Power Integrity Modeling of High-Speed Memory Modules Using Chip-Package-Board Coanalysis”, *IEEE Transactions on Energy Conversion* 52(2) (2010), pp. 381-391.
- [81] B. Bae, S. Kim, Y. Kim, S. Kang, I. J. Kim, K. Kim, S. Kang and K. J. Han, “A Preliminary Analysis of Domain Coupling in Package Power Distribution Network”, *Proc. RFIT*, 2017, pp. 19-21.
- [82] Y. Panov and M. M. Jovanovic, “Design Considerations for 12-V/1.5-V, 50-A Voltage Regulator Modules”, *IEEE Transactions on Power Electronics* 16(6) (2001), pp. 776-783.

- [83] E. H. K. Hsiung, Y. L. Li, R. B. Wu, T. Su, Y. S. Cheng and K. B. Wu, “A Linear 4-Element Model of VRM - Characteristics, Practical Uses and Limitations”, *Proc. EDAPS*, 2012, pp. 13-16.
- [84] K. L. Kaiser, *Electromagnetic compatibility handbook*, CRC press, 2004.
- [85] CPM, ANSYS, <http://www.ansys.com>
- [86] B. Ross, “IBIS Models for Signal Integrity Applications”, *EE Times* 18(9) (1996), pp. 38-43.
- [87] R. Stilkol, Apache Design Solutions, “Chip-Package-System (CPS) Co-Design Verification”, *Chipex* 2011.
- [88] J. Kim, S. Wu, H. Wang, Y. Takita, H. Takeuchi, K. Araki, G. Feng and J. Fan, “Improved Target Impedance and IC Transient Current Measurement for Power Distribution Network Design”, *Proc. EMC*, 2010, pp. 445-450.
- [89] Y. E. Chen, T. H. Tsai, S. H. Chen and H. M. Chen, “Cost-effective Decap Selection for Beyond Die Power Integrity”, *Proc. DATE*, 2014, pp. 1-4.
- [90] W. H. Hayt and J. A. Buck, *Engineering electromagnetics*, McGraw-Hill, 1981.
- [91] R. C. Paul and G. A. Macon, “What is Partial Inductance”, *Proc. EMC*, 2008, pp. 1-23.
- [92] J. D. Jackson, *Classical Electrodynamics*, Wiley, 1975.
- [93] Linley group, <http://www.economist.com>
- [94] Intel, 2018,
<https://newsroom.intel.com/news/new-intel-architectures-technologies-target-expanded-market-opportunities>
- [95] Intel, 2018,
<https://www.extremetech.com/extreme/282950-beyond-traditional-computing-expectations-for-intel-in-2019>
- [96] ITRS, 2012, <http://public.itrs.net>
- [97] HSPICE, 2013, Synopsys, version K-2013.03-SP1, <http://www.synopsys.com>
- [98] HSPICE, 2015, Synopsys, version K-2015.06-2, <http://www.synopsys.com>
- [99] HFSS, 2013, ANSYS, version 16.2, <http://www.ansys.com>
- [100] 22 nm PTM HP model, <http://ptm.asu.edu>